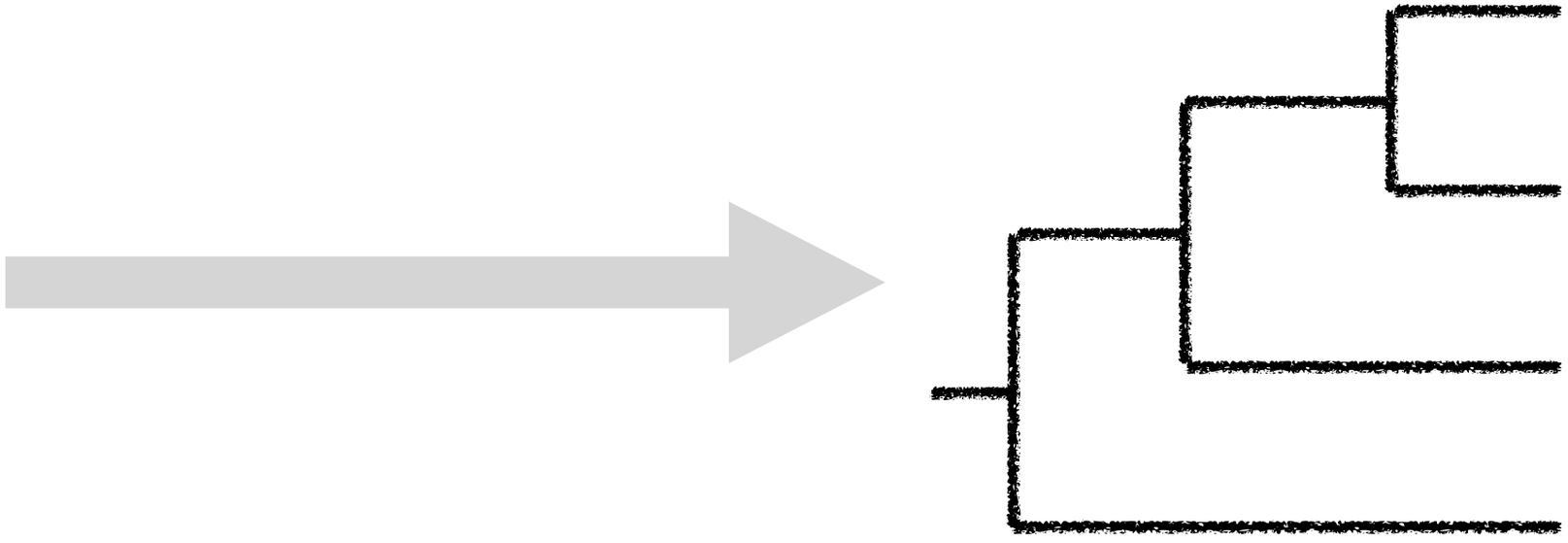
Lecture 14

Coalescent-based methods
Botany 563 — Spring 2021

- **Previous class check-up:**
    - We studied Bayesian phylogenetic inference
    - We practiced on MrBayes and/or PhyloBayes
- **Learning Objectives:** At the end of today's session, you will be able to
    - Explain the coalescent model on a species tree
    - Explain the coalescent model on a species network
    - Explain the steps in coalescent-based methods and the comparison with concatenation approaches
- **Pre-class work**
    - Read HAL 3.1 and 3.3

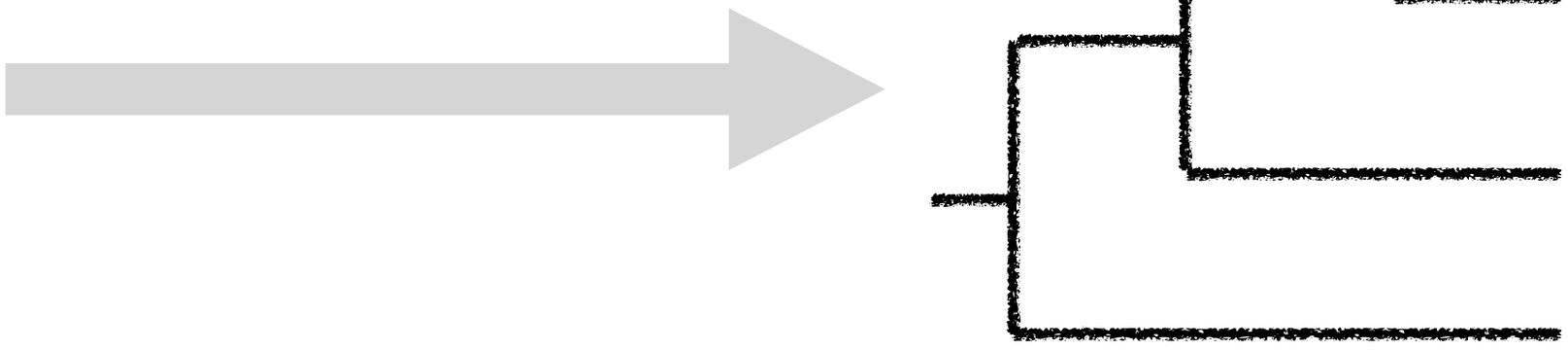
Phylogenetic inference

AAGTCTAG
AAGTCTAG
AACTCTAG
AATTCTAG



Phylogenetic inference

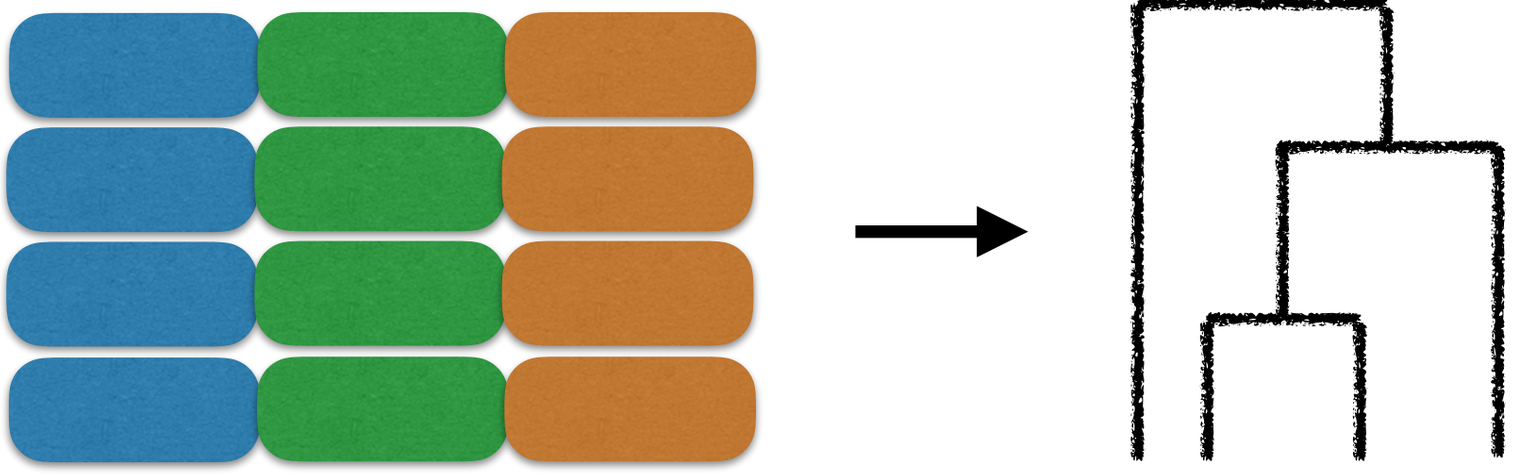
AAGTCTAG
AAGTCTAG
AACTCTAG
AATTCTAG



Gene
Locus
Region
Whole genome

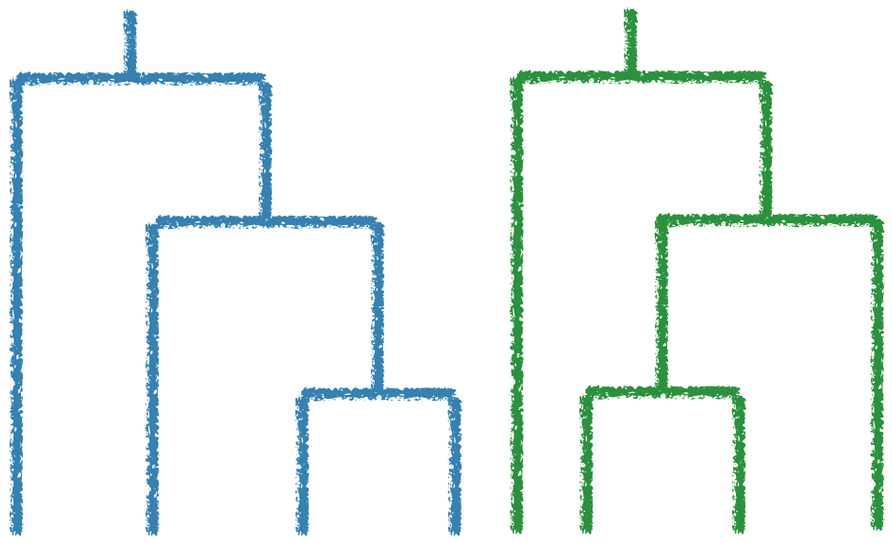
Statistically inconsistent

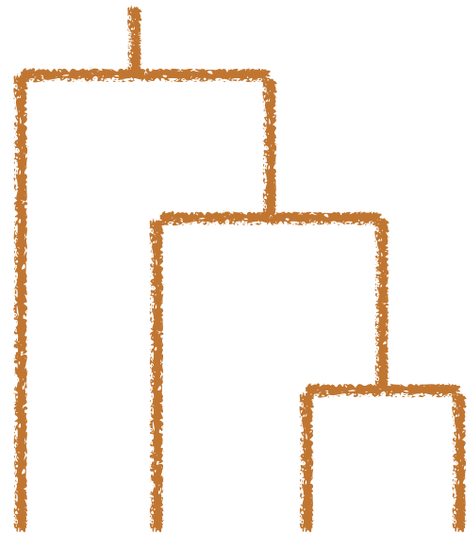
(Kubatko, Degnan, 2007)
(Roch, Steel, 2015)



Concatenation or supermatrix

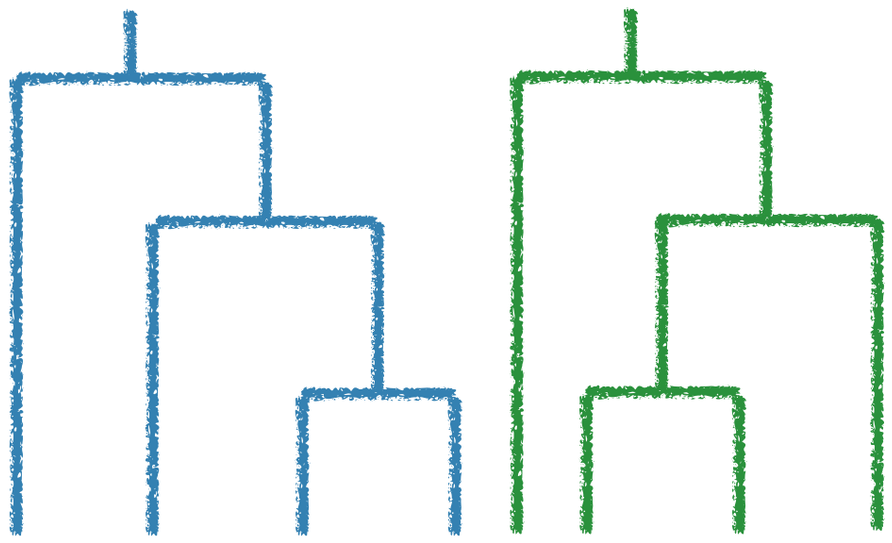
MrBayes
(Huelsenbeck, Ronquist, 2001)
RAxML
(Stamatakis, 2014)
IQ-tree 2
(Minh et al, 2020)

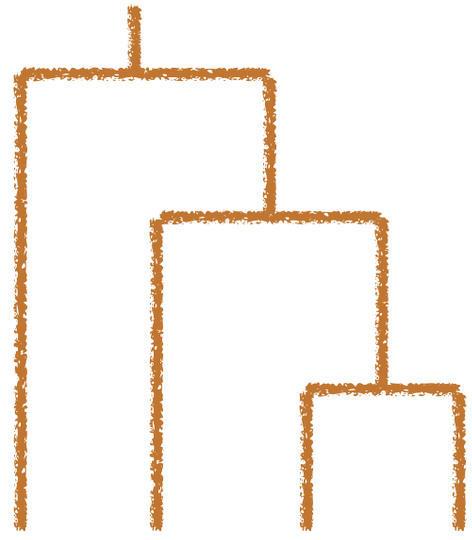


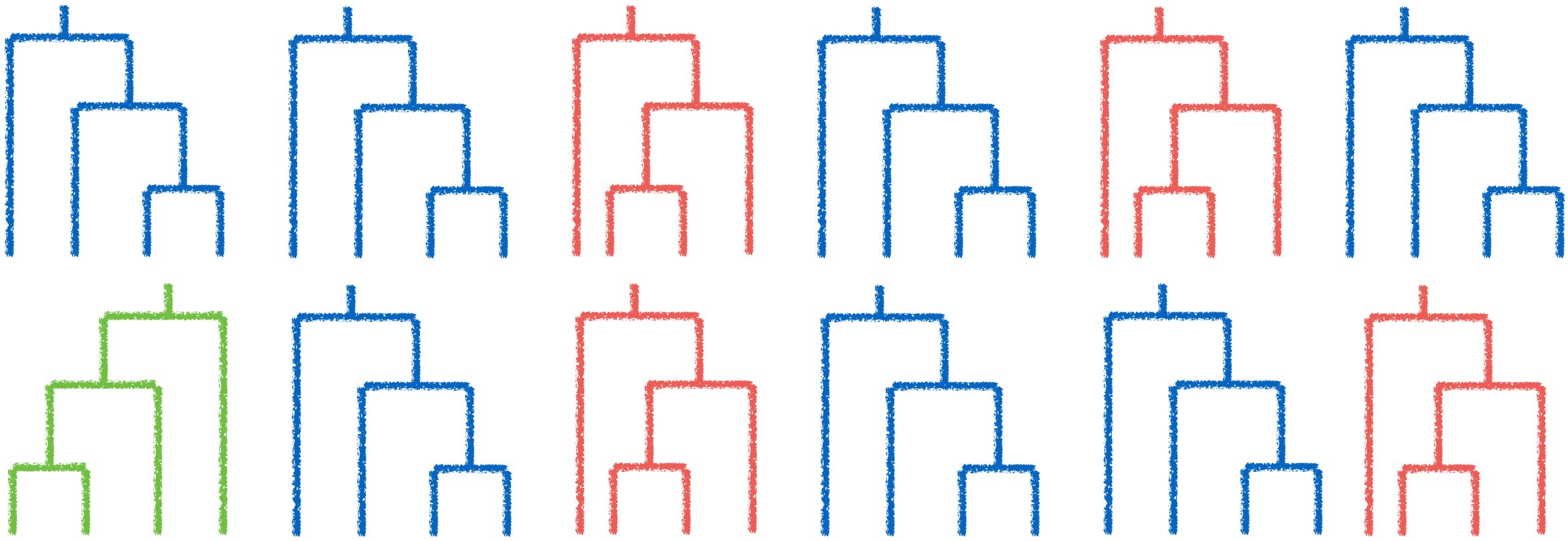


Estimate gene trees

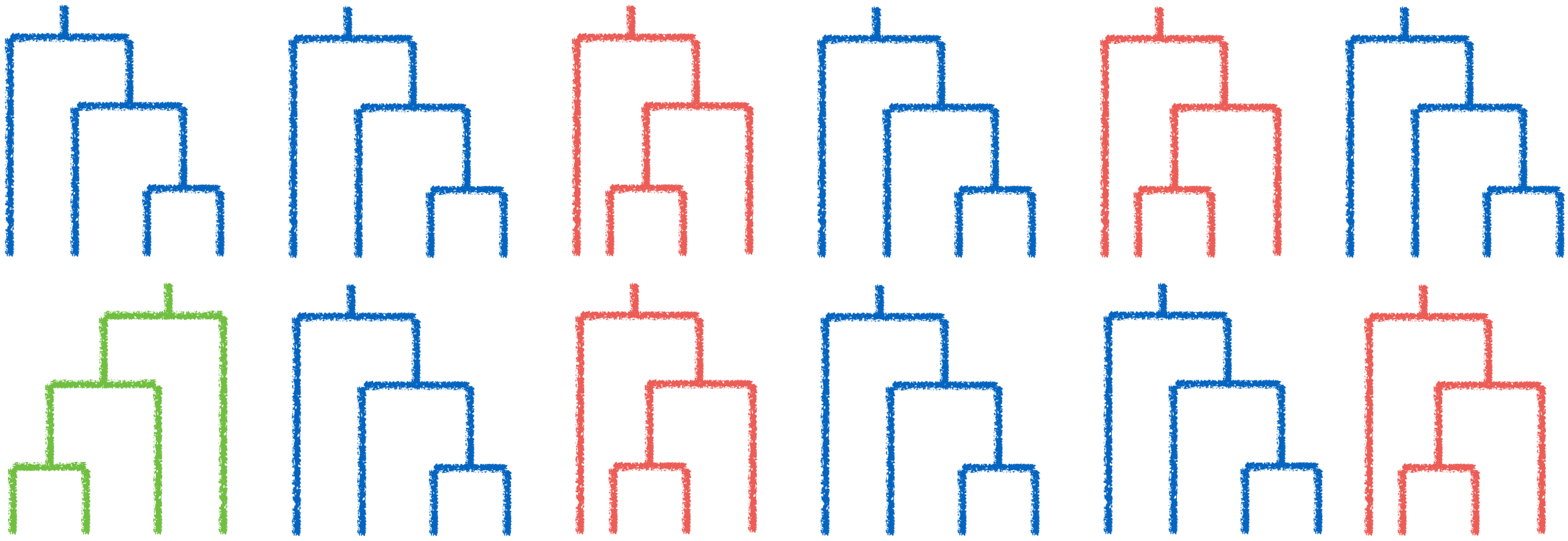
MrBayes
(Huelsenbeck, Ronquist, 2001)
RAxML
(Stamatakis, 2014)
IQ-tree 2
(Minh et al, 2020)





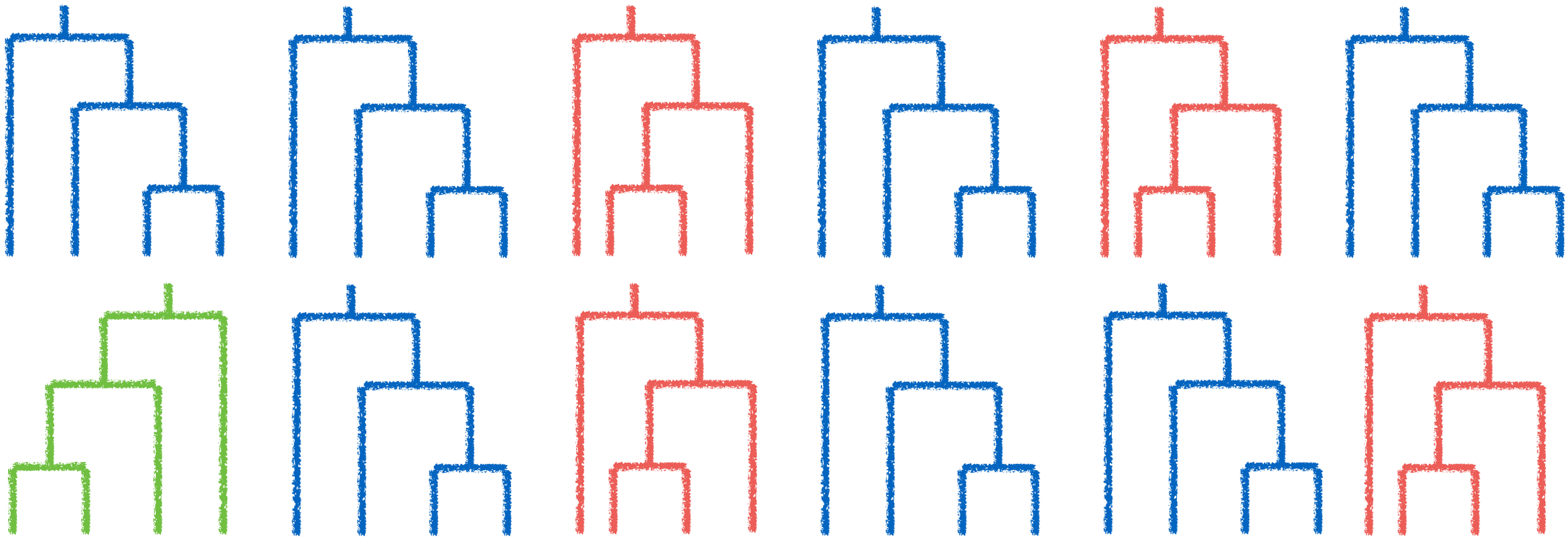


Gene trees



Gene trees

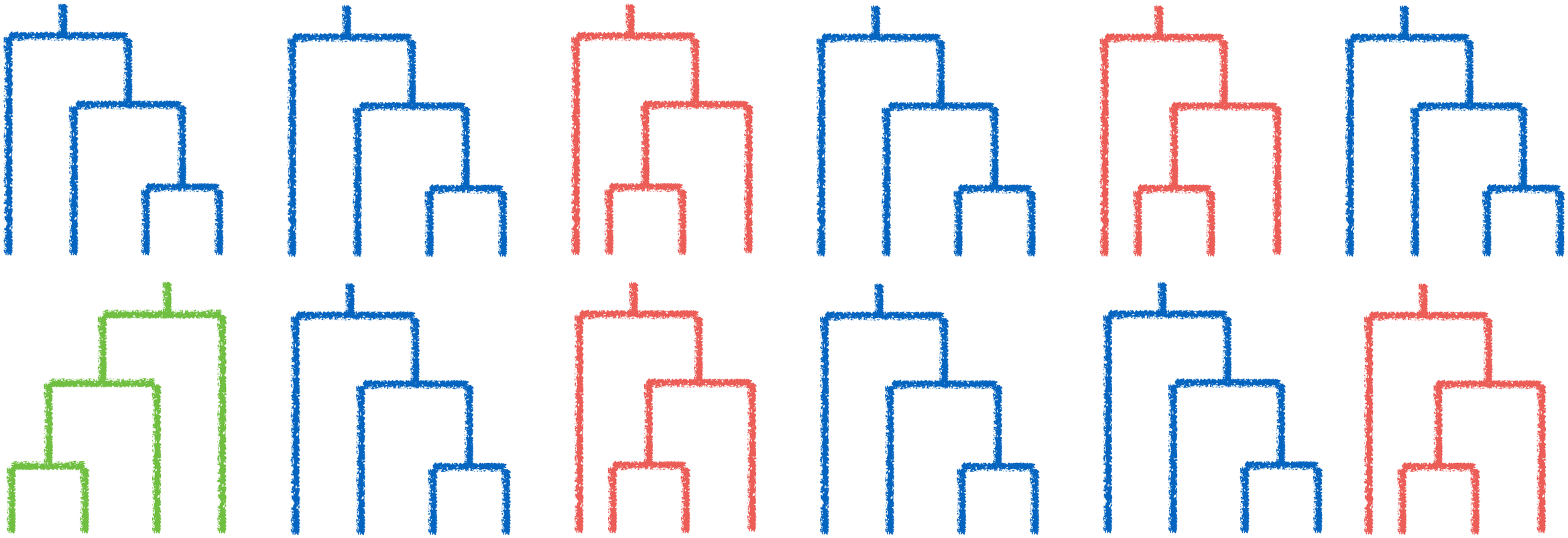
- Estimation error
- Incomplete lineage sorting
- Gene flow



Gene trees

- Estimation error
- Incomplete lineage sorting
- Gene flow

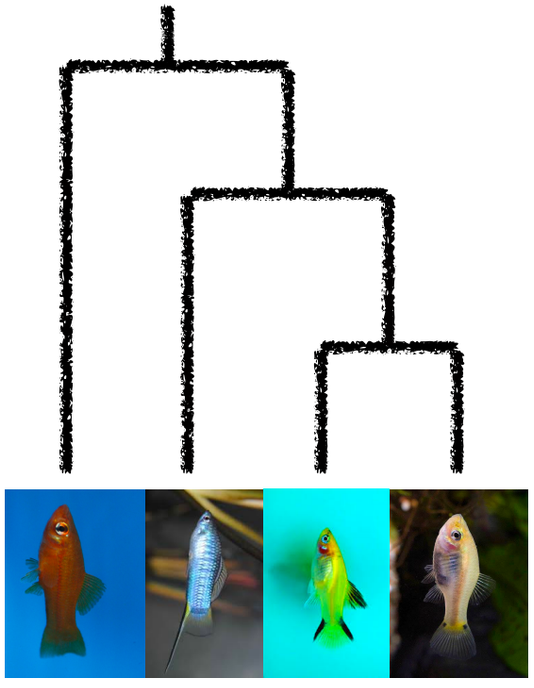
Species evolutionary history



Gene trees

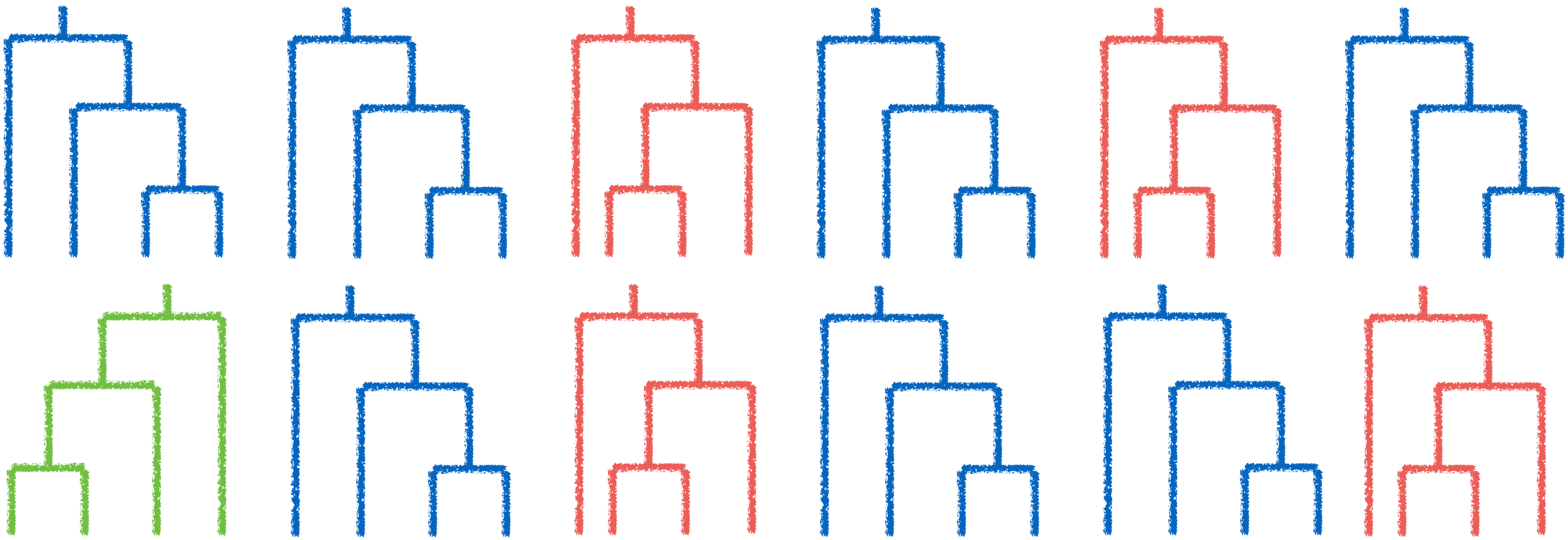
- Estimation error
- Incomplete lineage sorting
- Gene flow

Species evolutionary history



BUCKy (Ané et al, 2007)

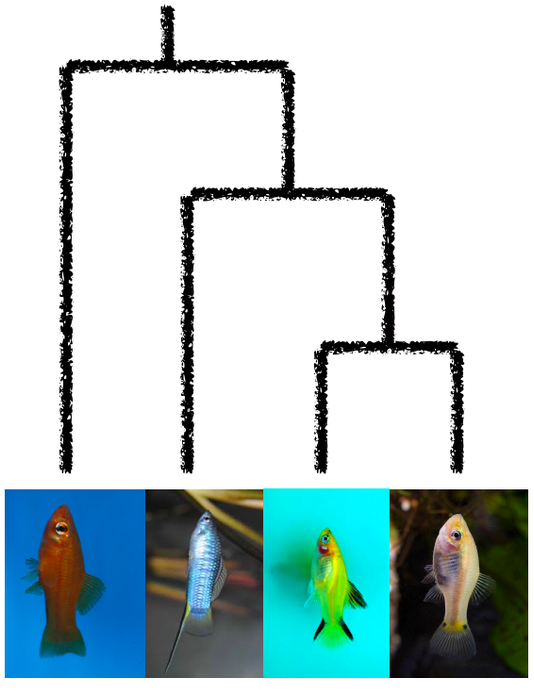
ASTRAL (Mirarab et al, 2014)



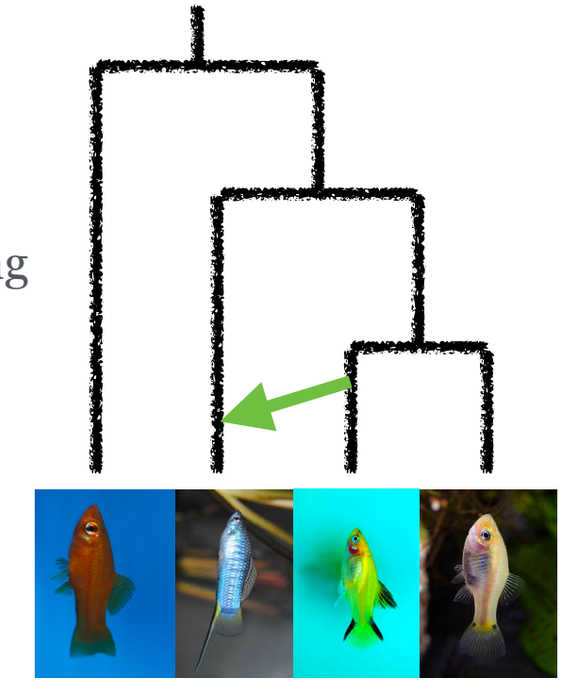
Gene trees

- Estimation error
- Incomplete lineage sorting
- Gene flow

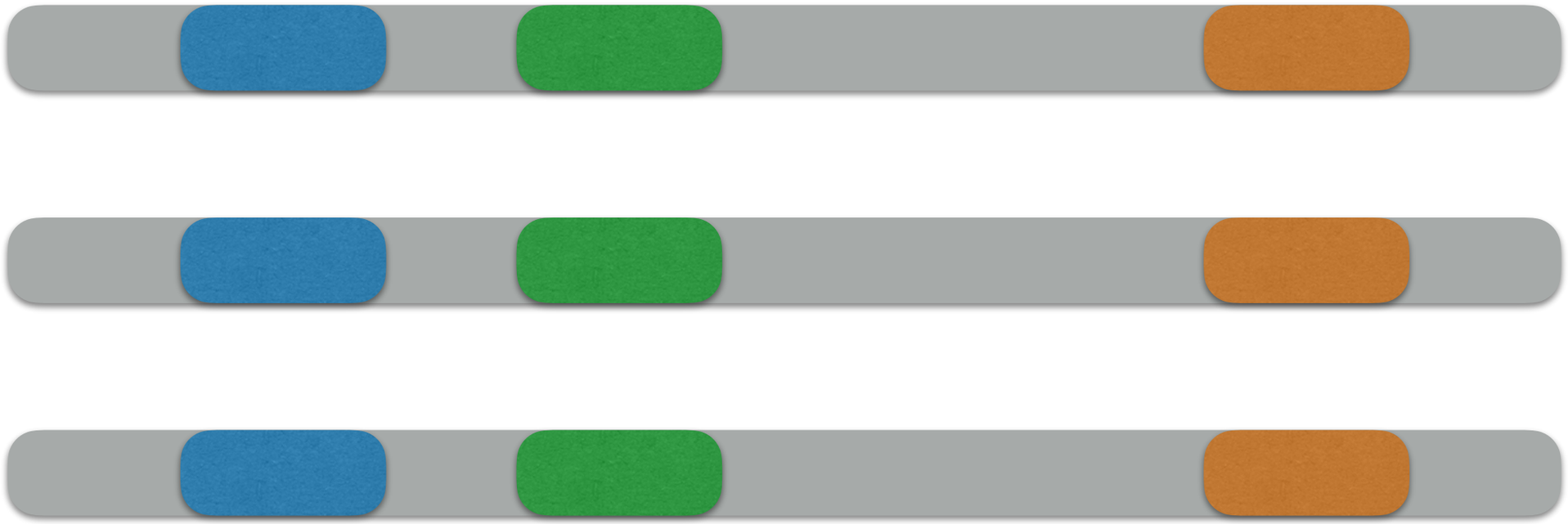
Species evolutionary history

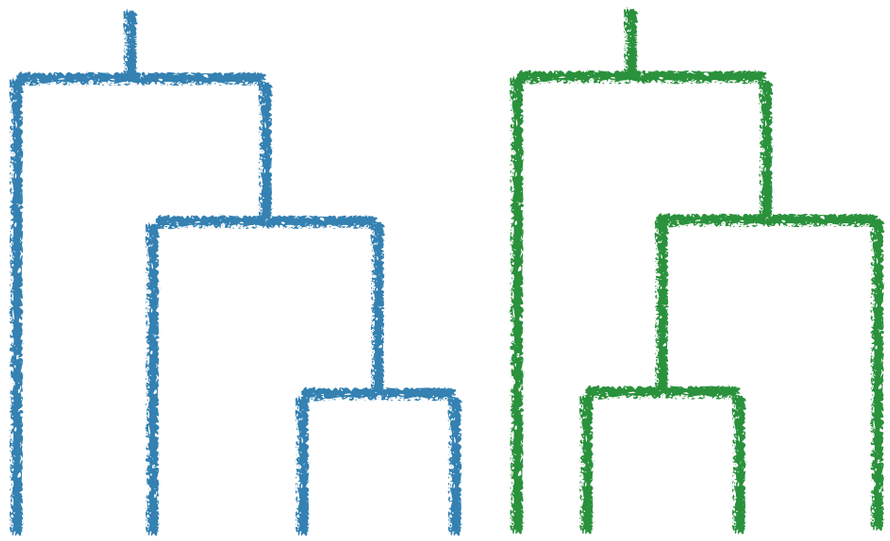


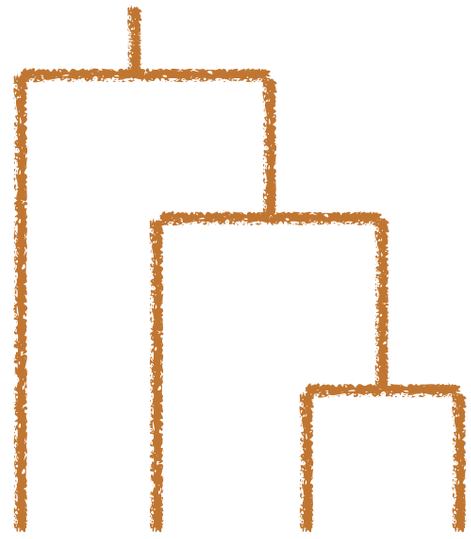
BUCKy (Ané et al, 2007)
ASTRAL (Mirarab et al, 2014)



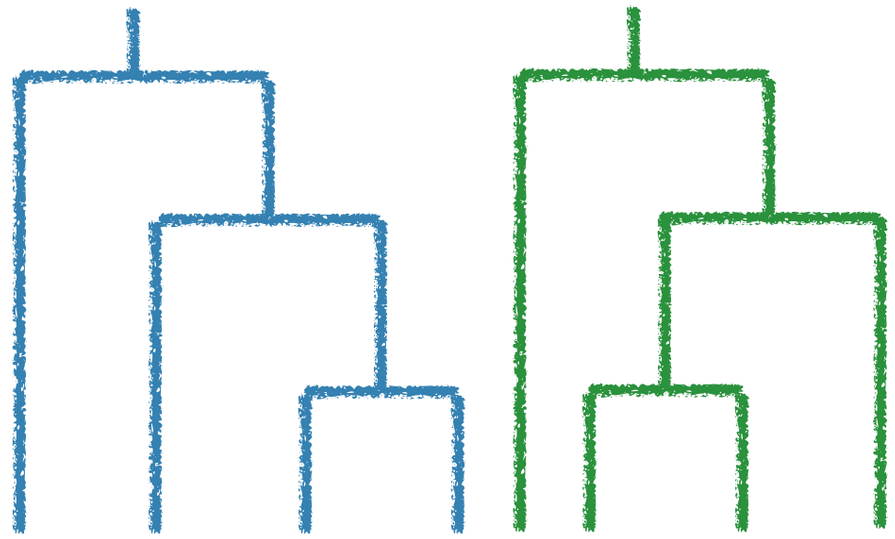
SNaQ
(Solís-Lemus, Ané, 2016)

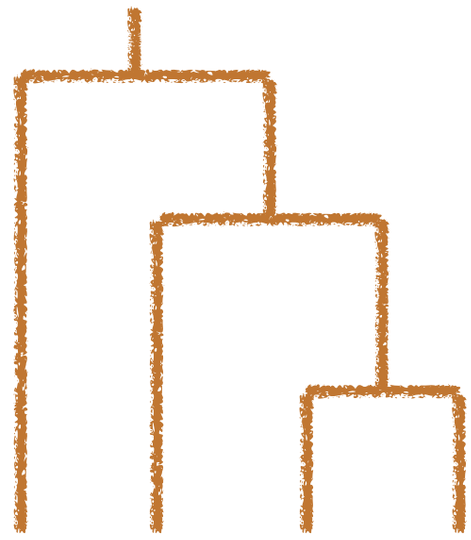




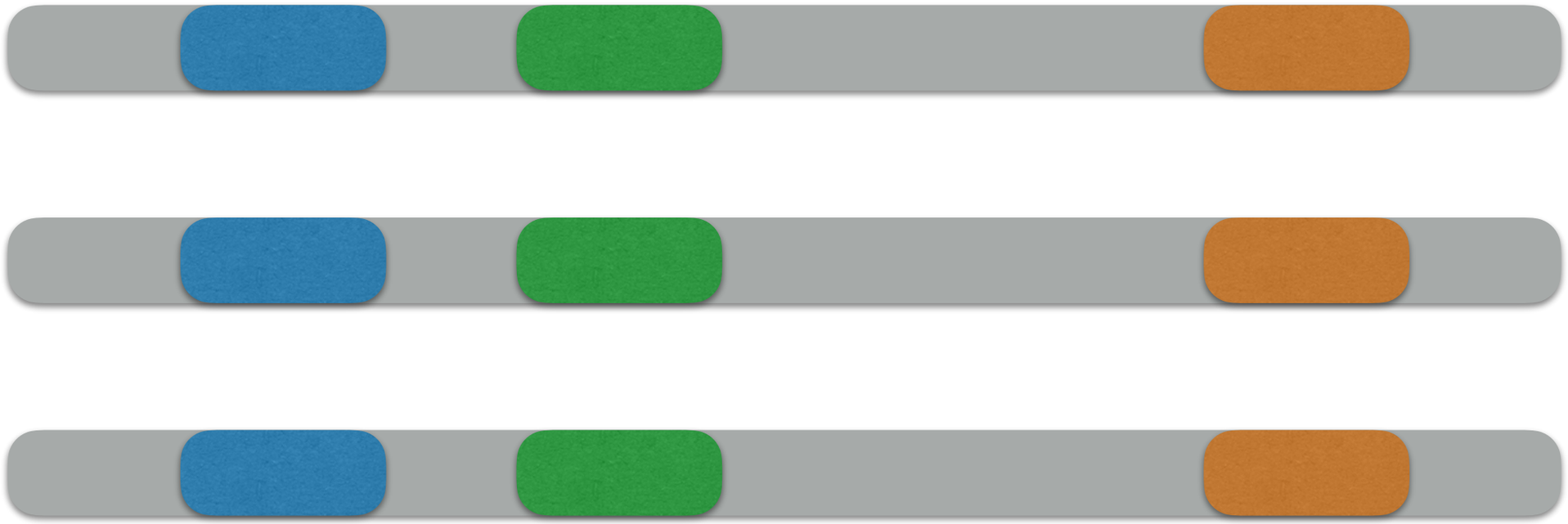


Ortholog
Recombination-free
MDL (Ané, 2011, GBE)



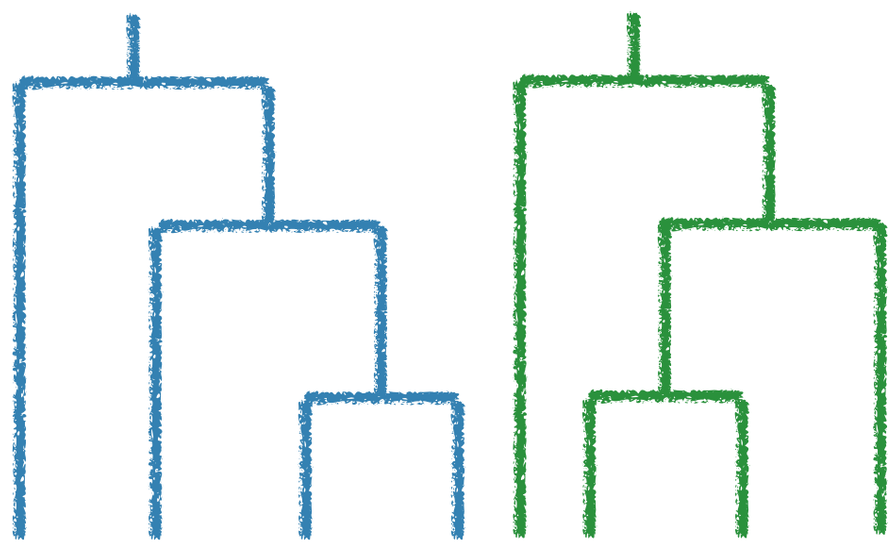


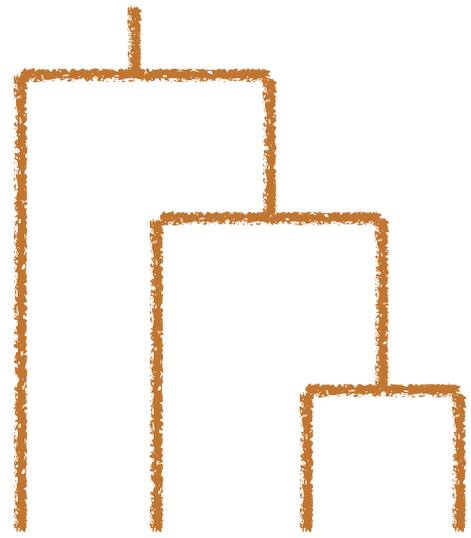
Ortholog
Recombination-free
MDL (Ané, 2011, GBE)



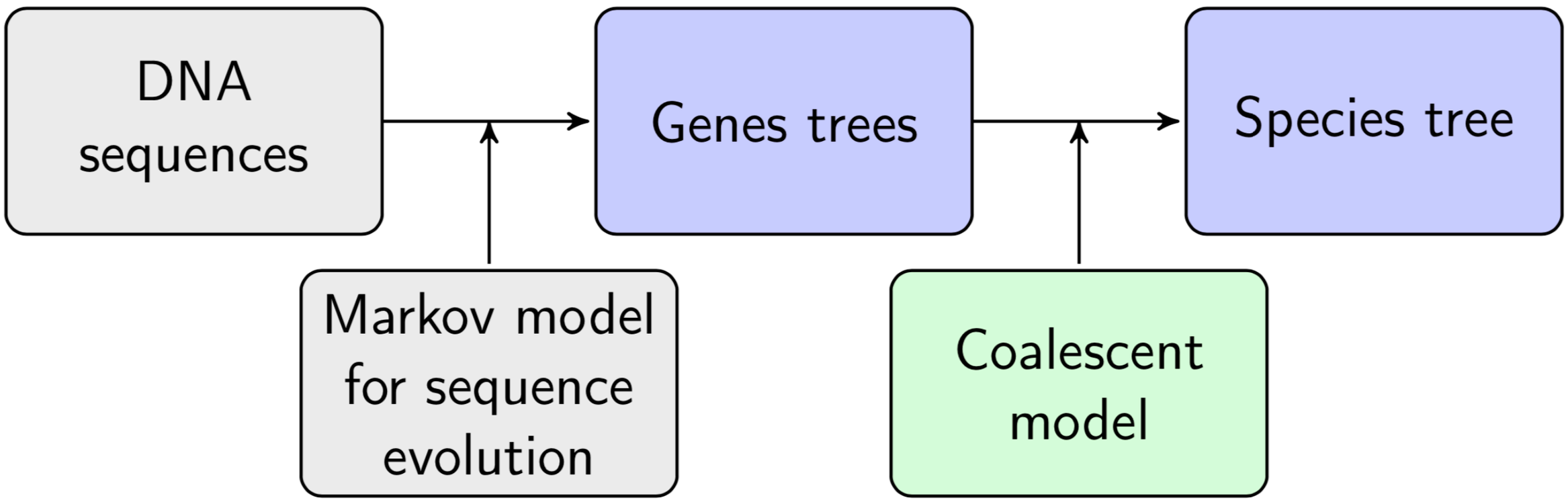
Estimate gene trees

MrBayes
(Huelsenbeck, Ronquist, 2001)
RAxML
(Stamatakis, 2014)
IQ-tree 2
(Minh et al, 2020)

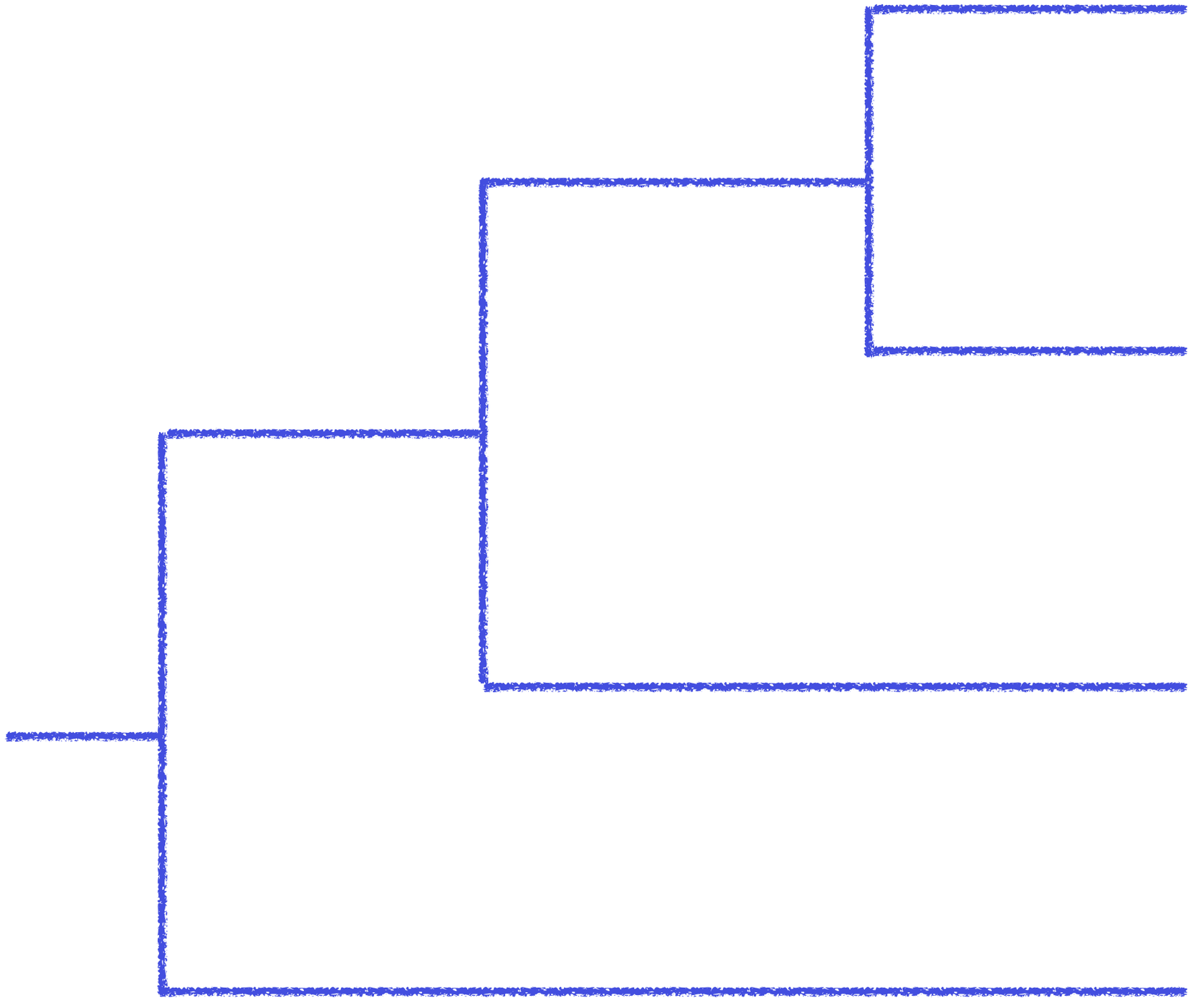




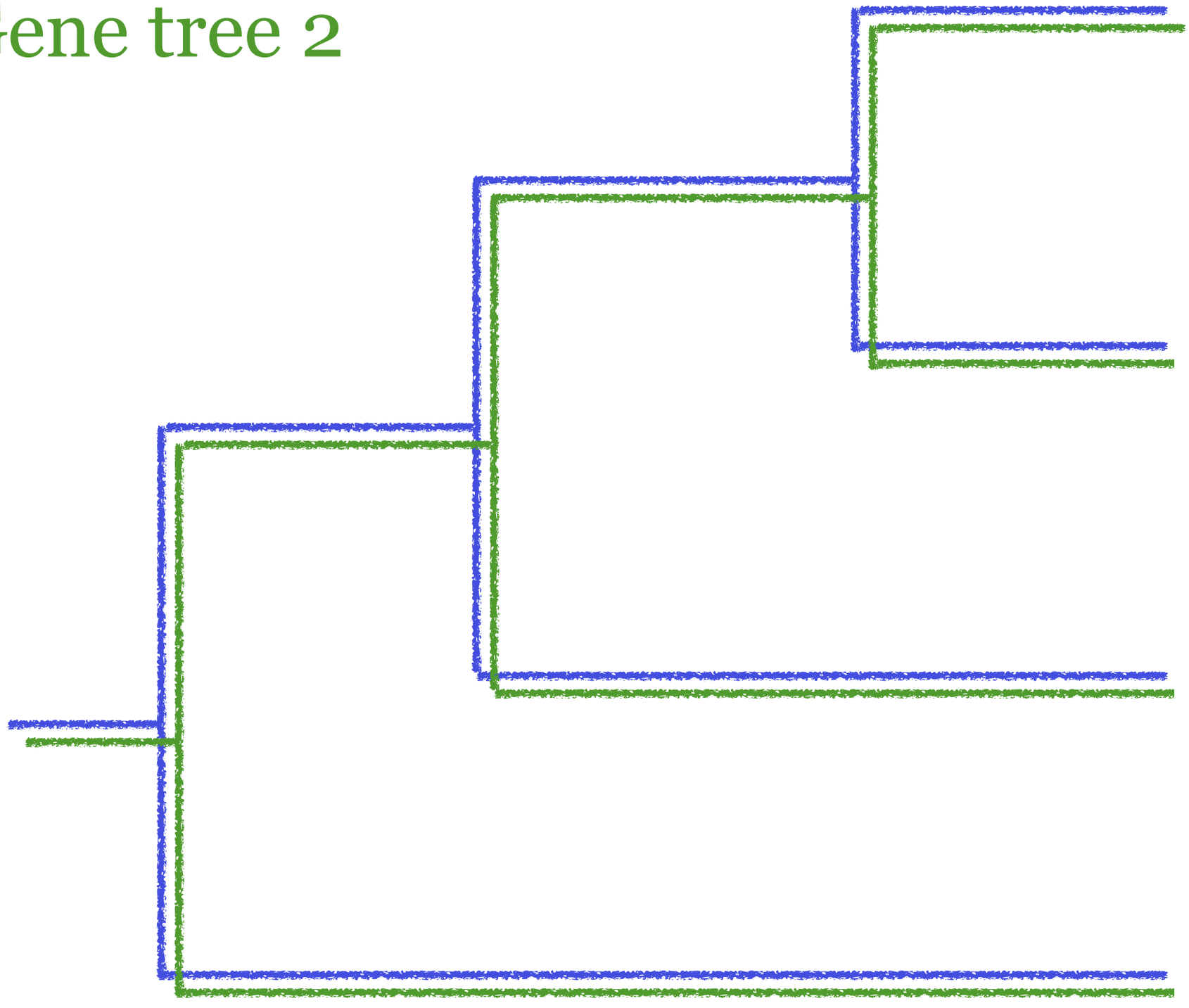
Phylogenetic inference



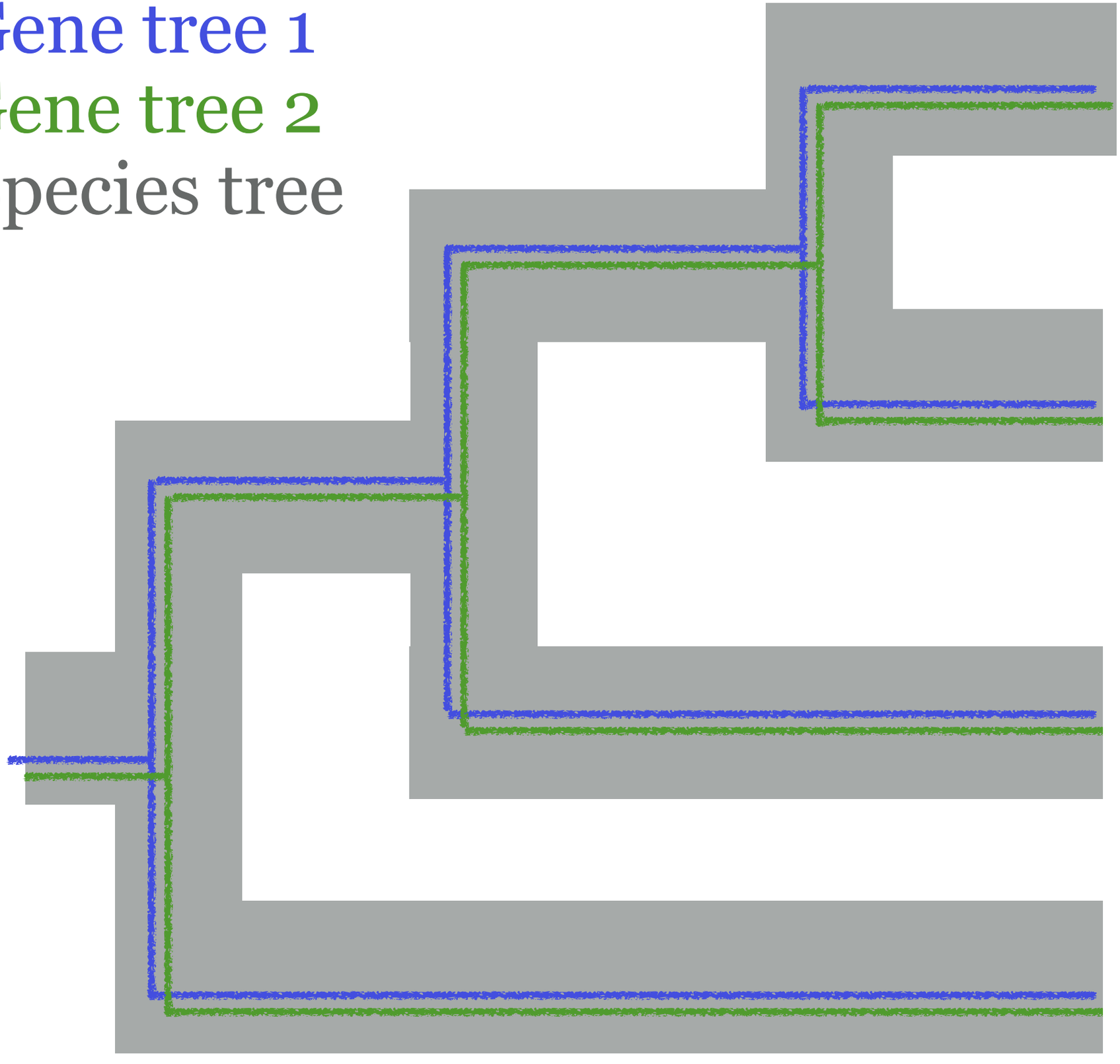
Gene tree



Gene tree 1
Gene tree 2



Gene tree 1
Gene tree 2
Species tree



Discordant gene tree

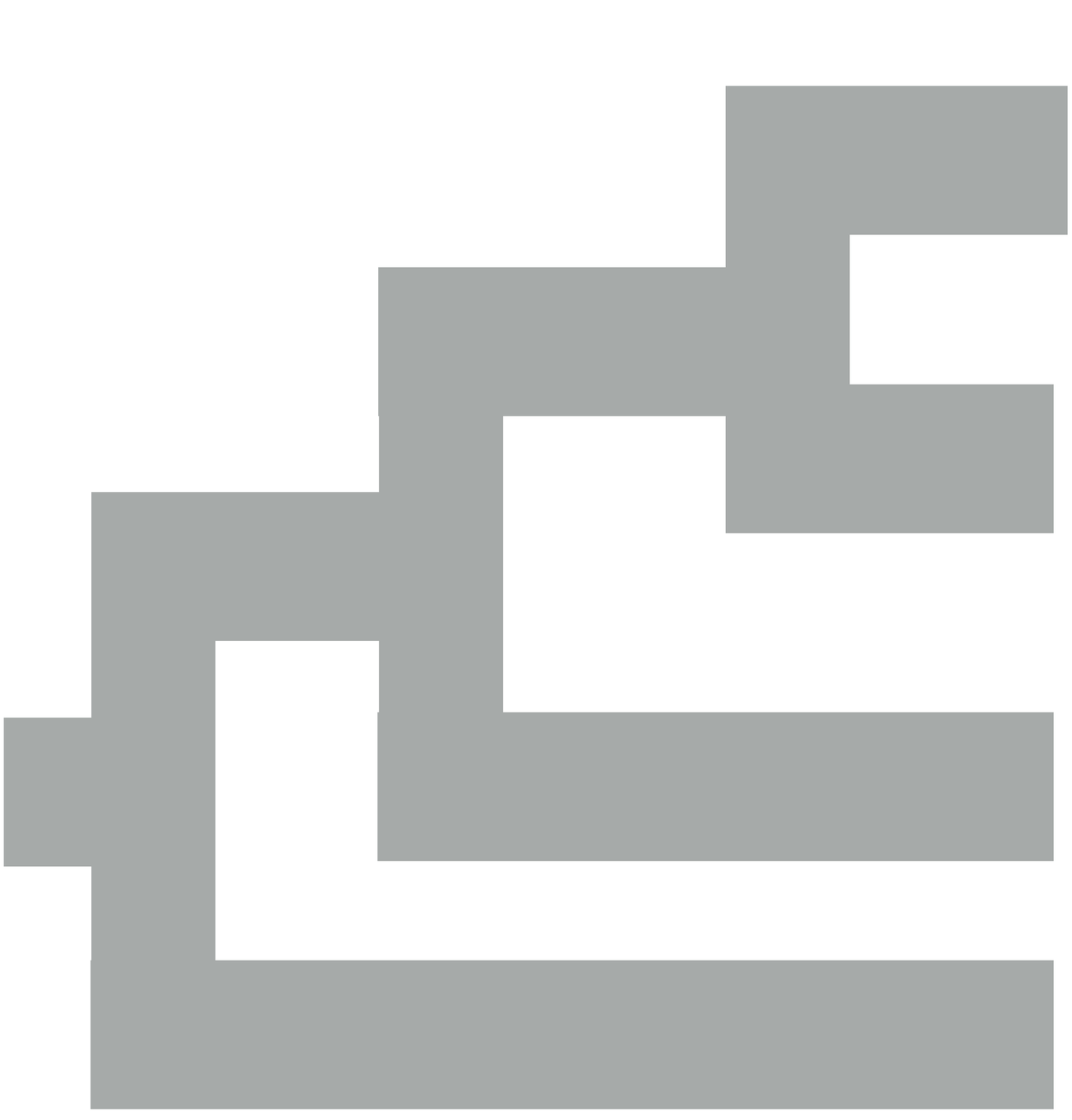
Species tree

Discordant gene tree

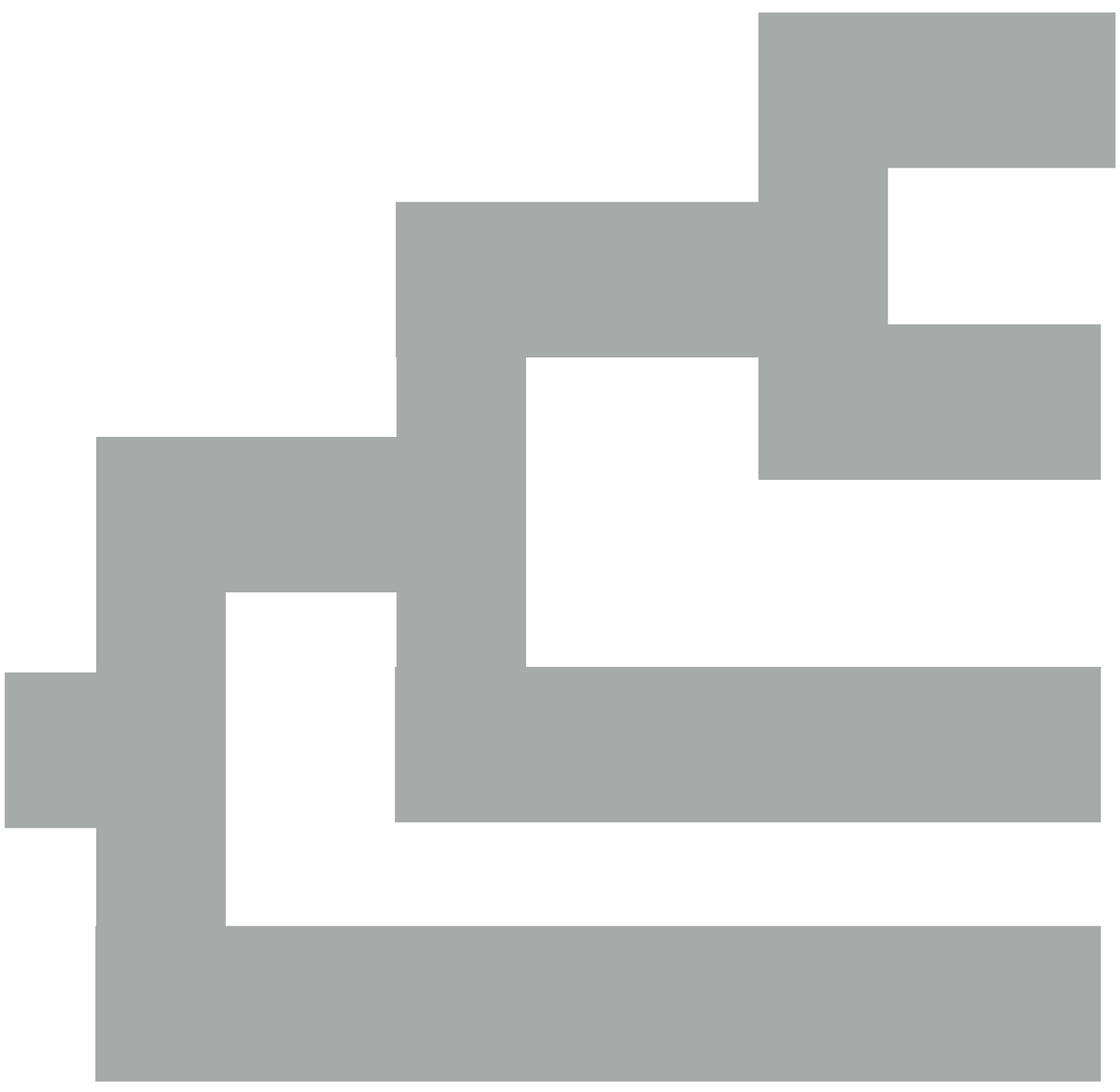
Species tree

~24%

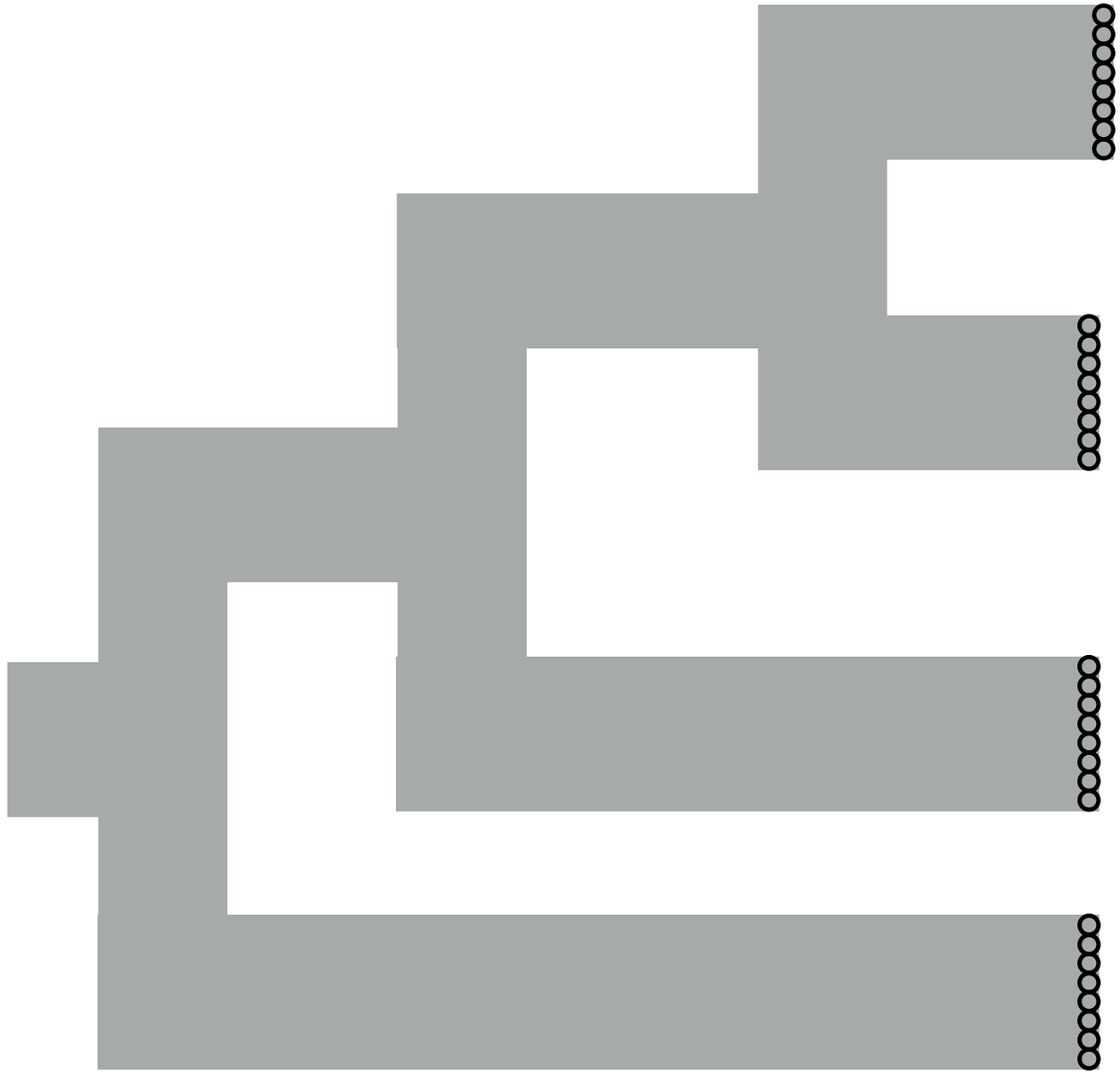
Coalescent model

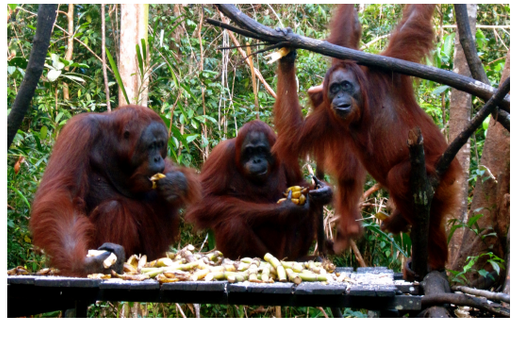


Coalescent model

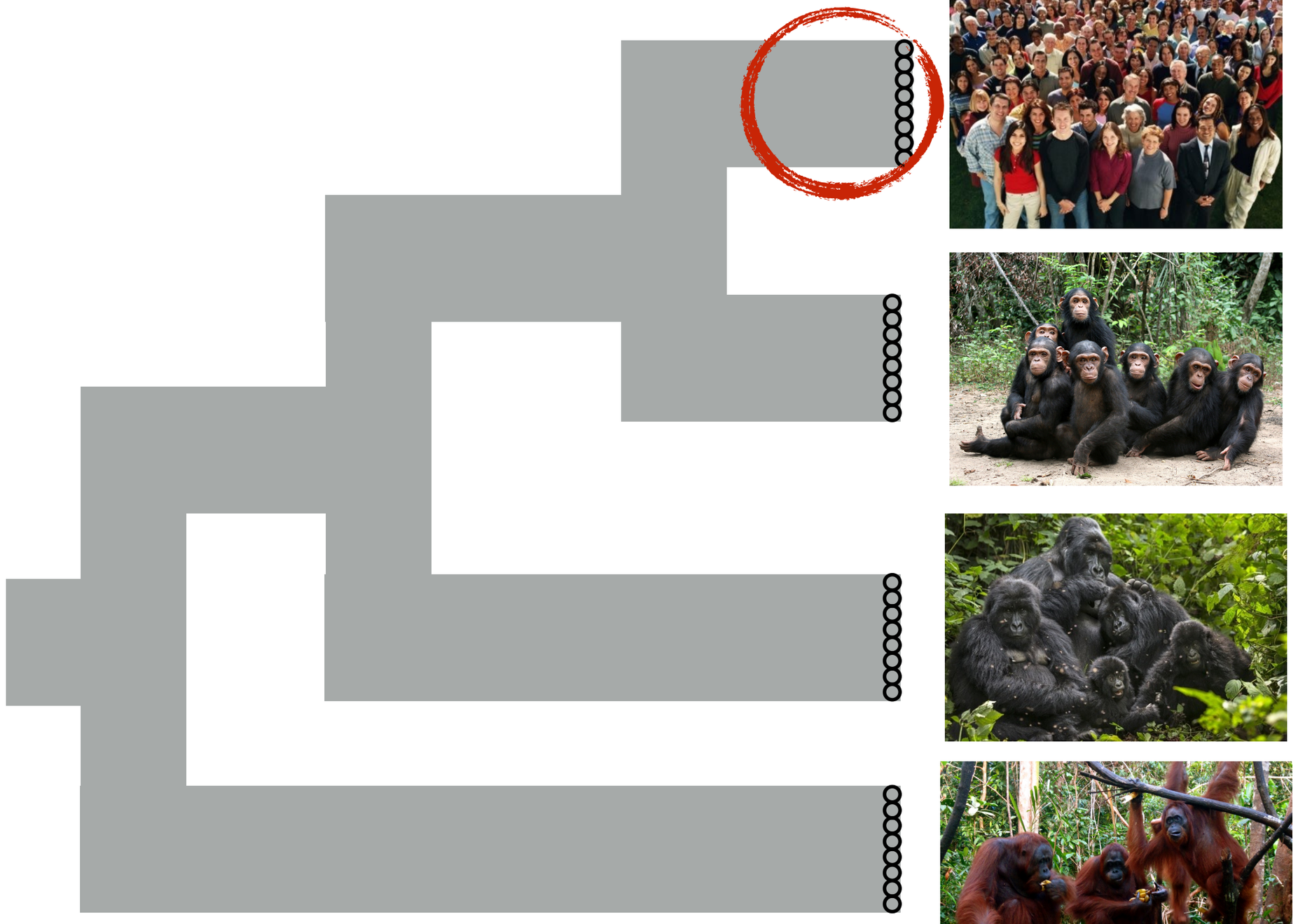


Coalescent model

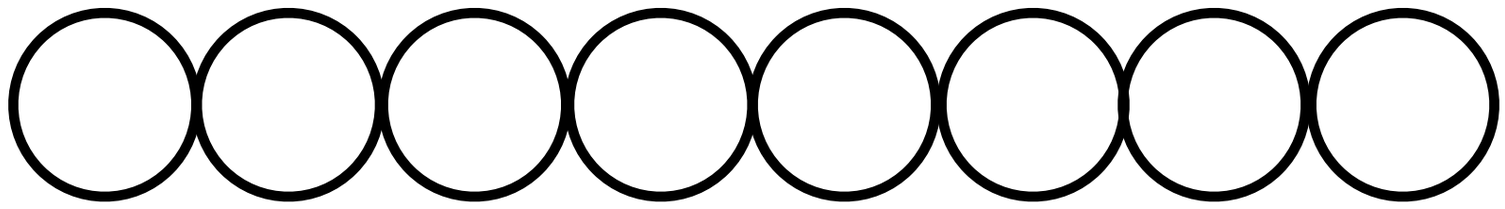




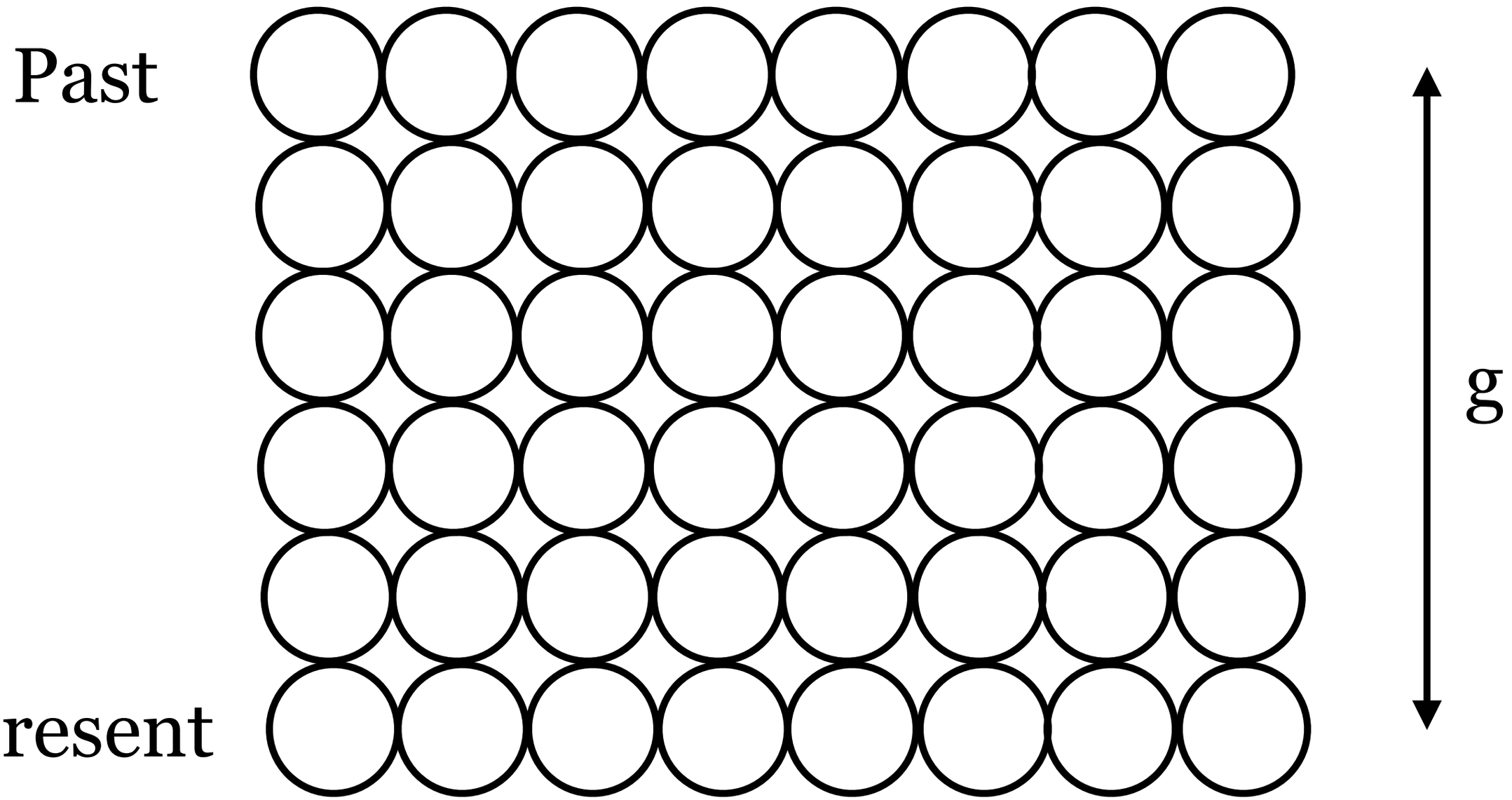
Coalescent model



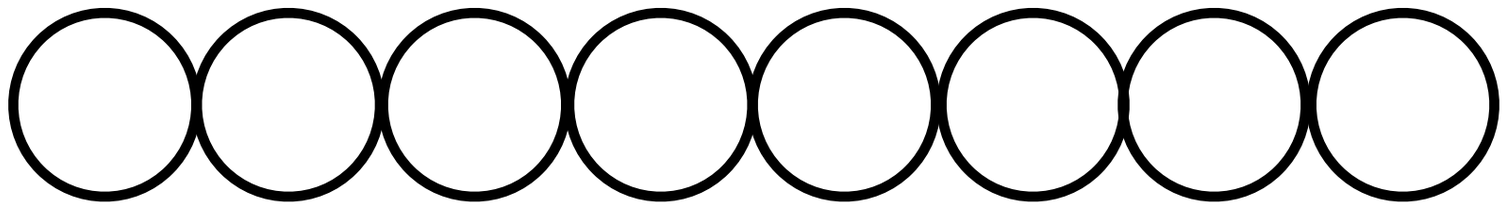
Coalescent model within 1 population

Present 

Coalescent model within 1 population

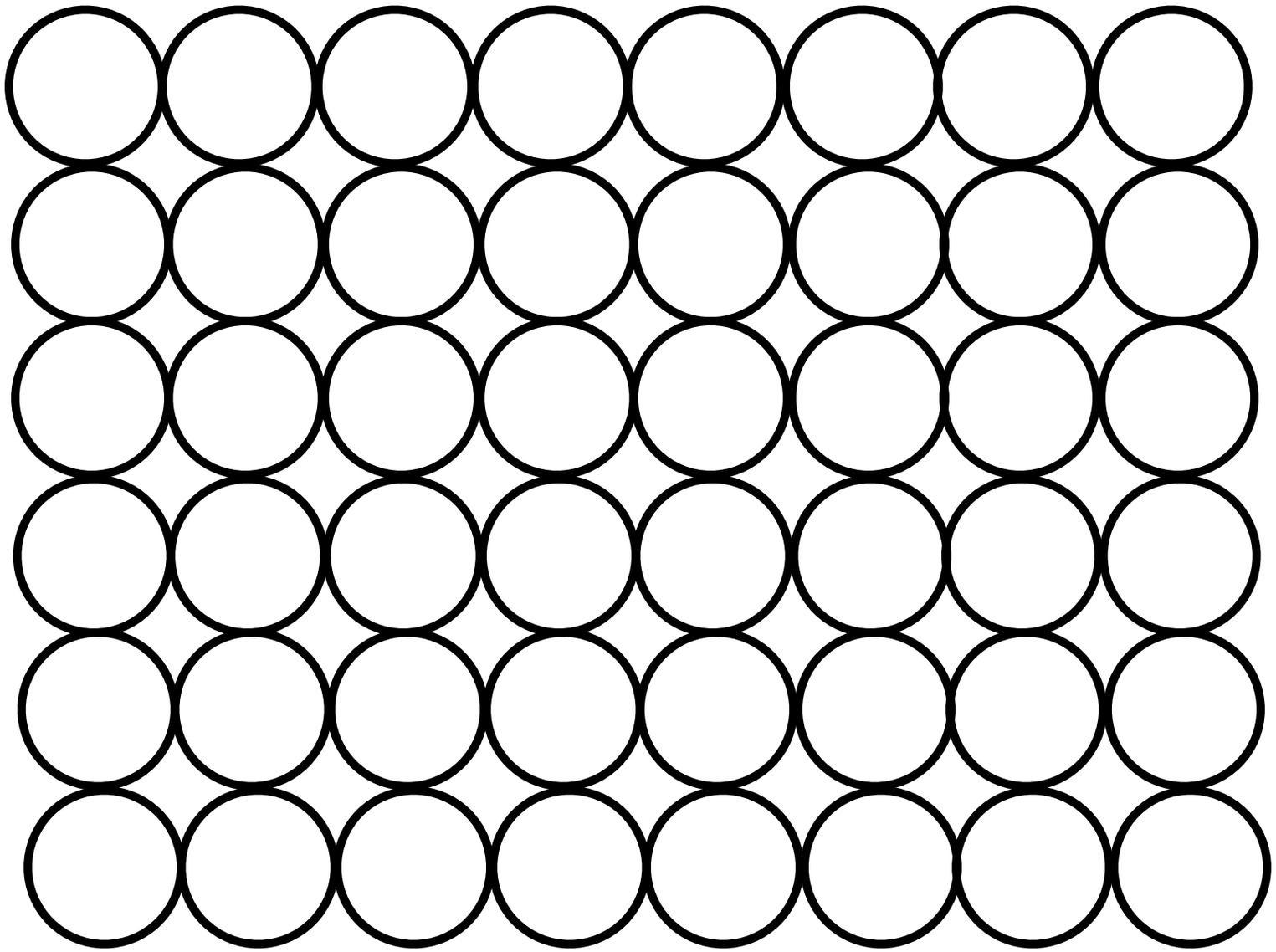


Coalescent model within 1 population

Present 

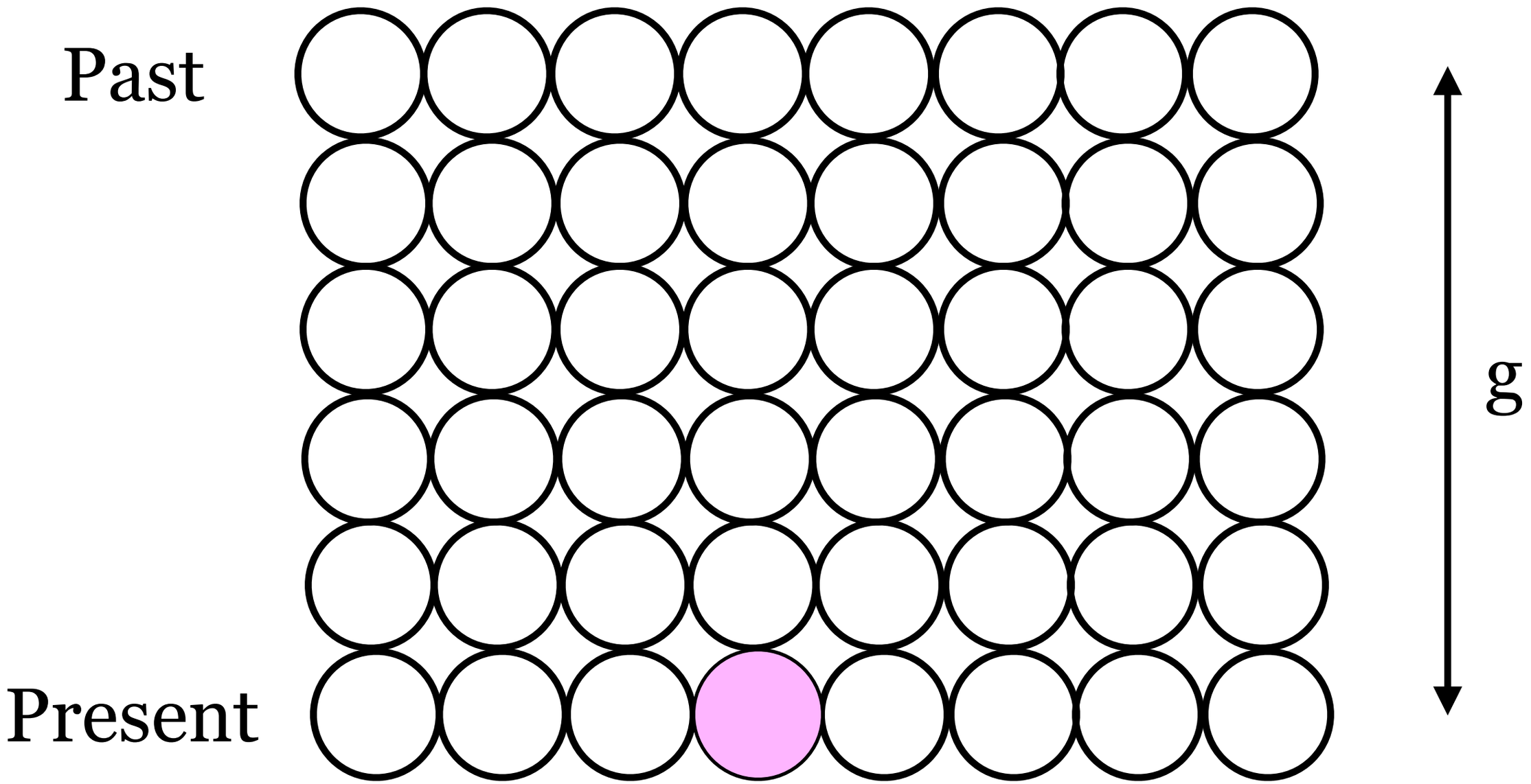
Coalescent model within 1 population

Past

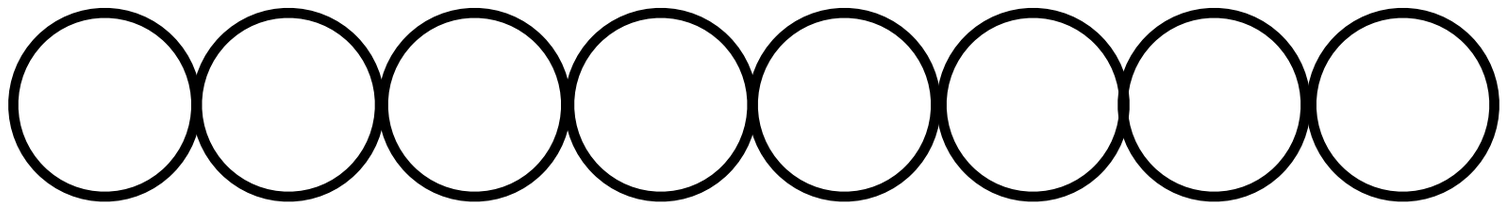


Present

Coalescent model within 1 population



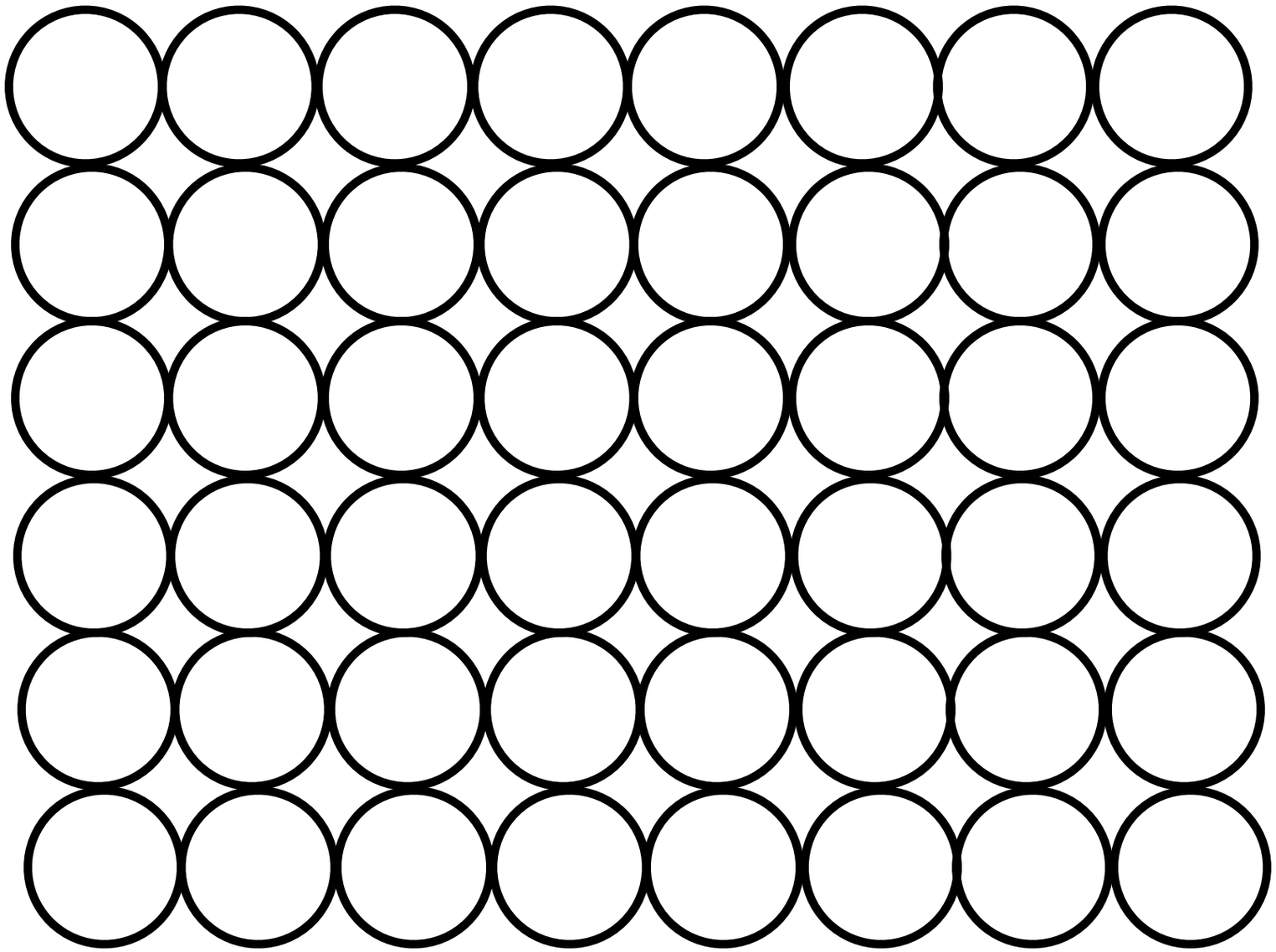
Coalescent model within 1 population

Present 

No selection: uniform probability 1/N

Coalescent model within 1 population

Past

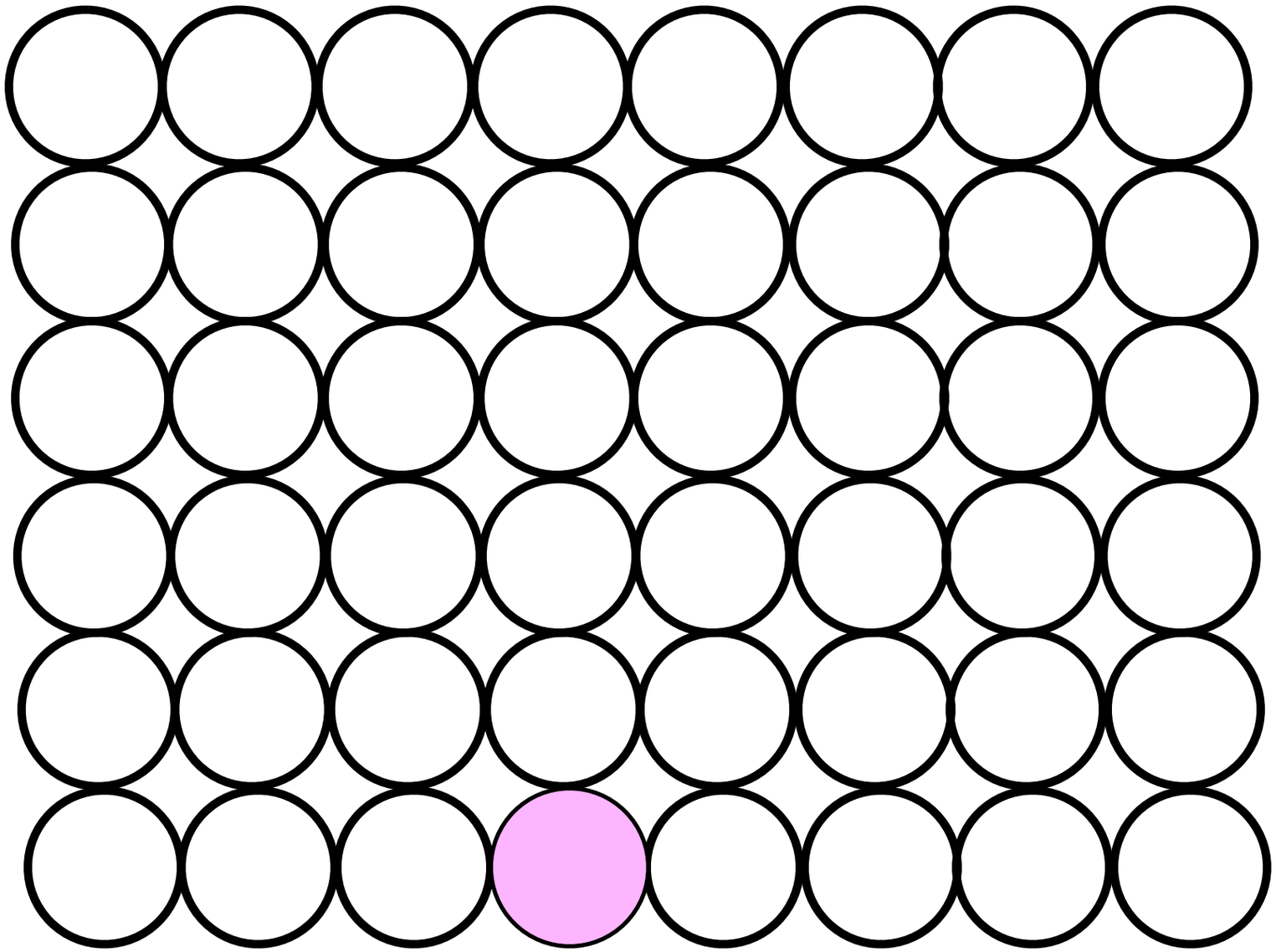


Present

No selection: uniform probability 1/N

Coalescent model within 1 population

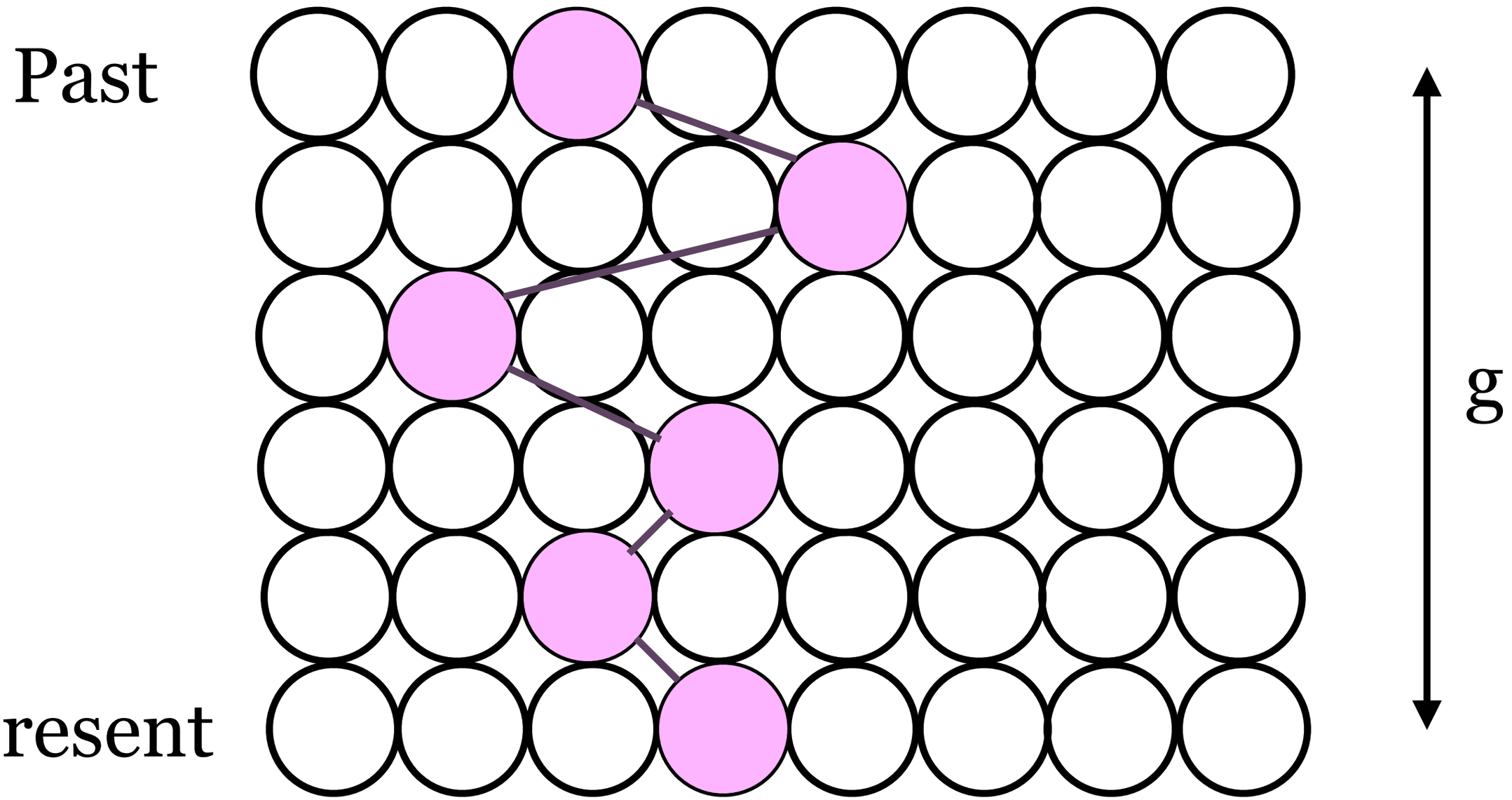
Past



Present

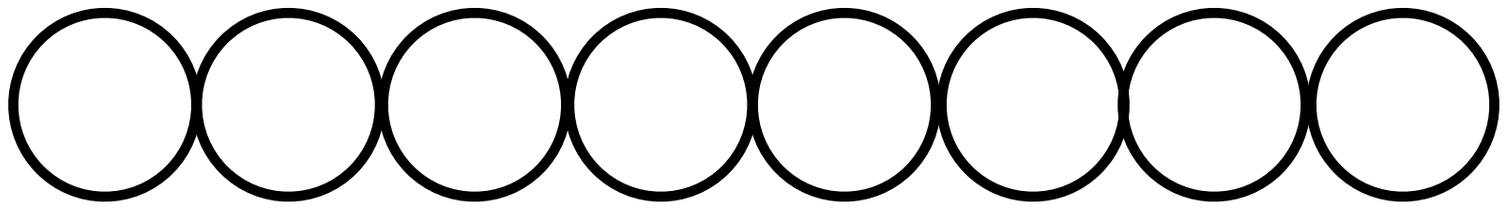
No selection: uniform probability 1/N

Coalescent model within 1 population



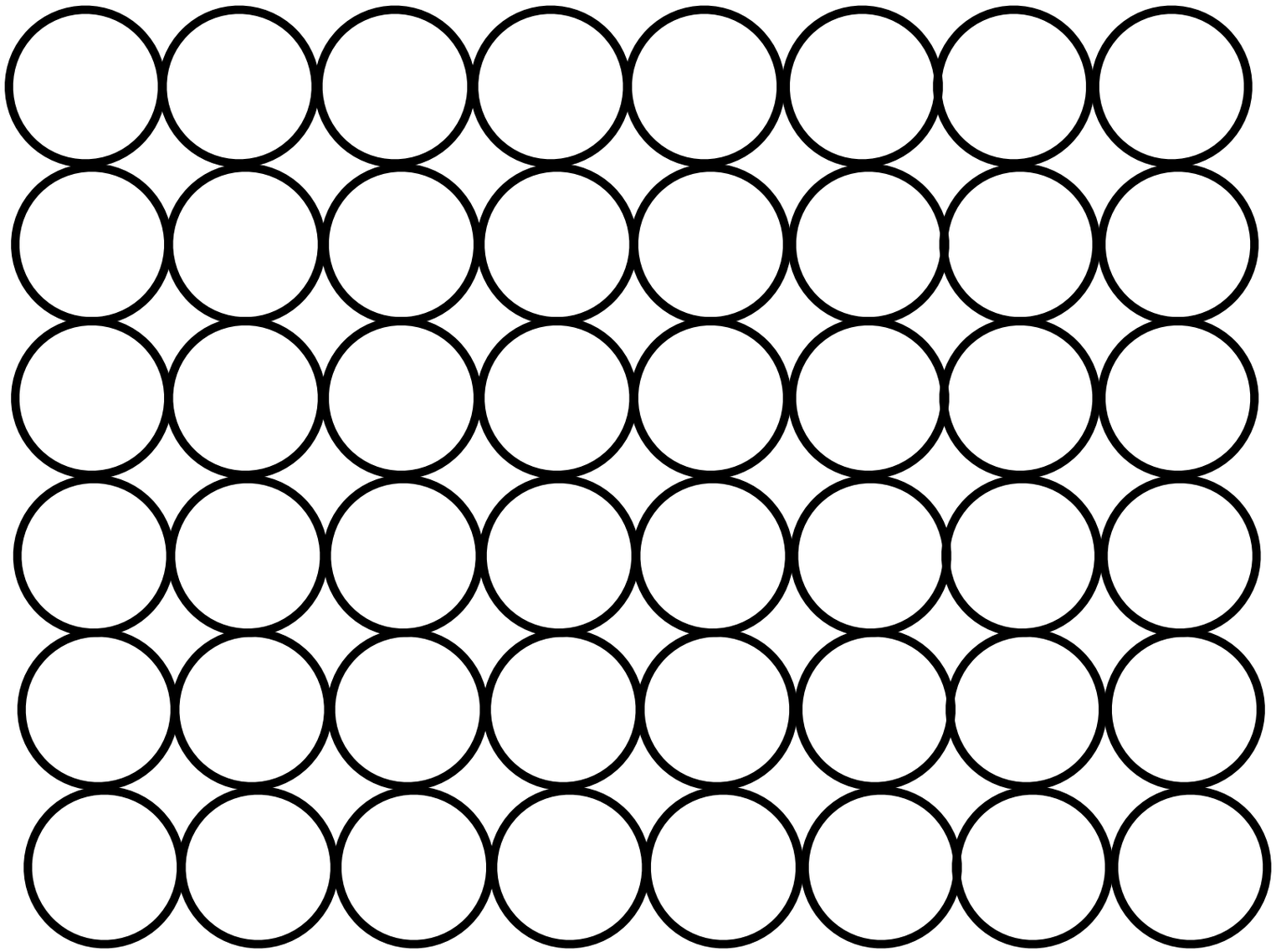
No selection: uniform probability 1/N

Coalescent model within 1 population

Present 

Coalescent model within 1 population

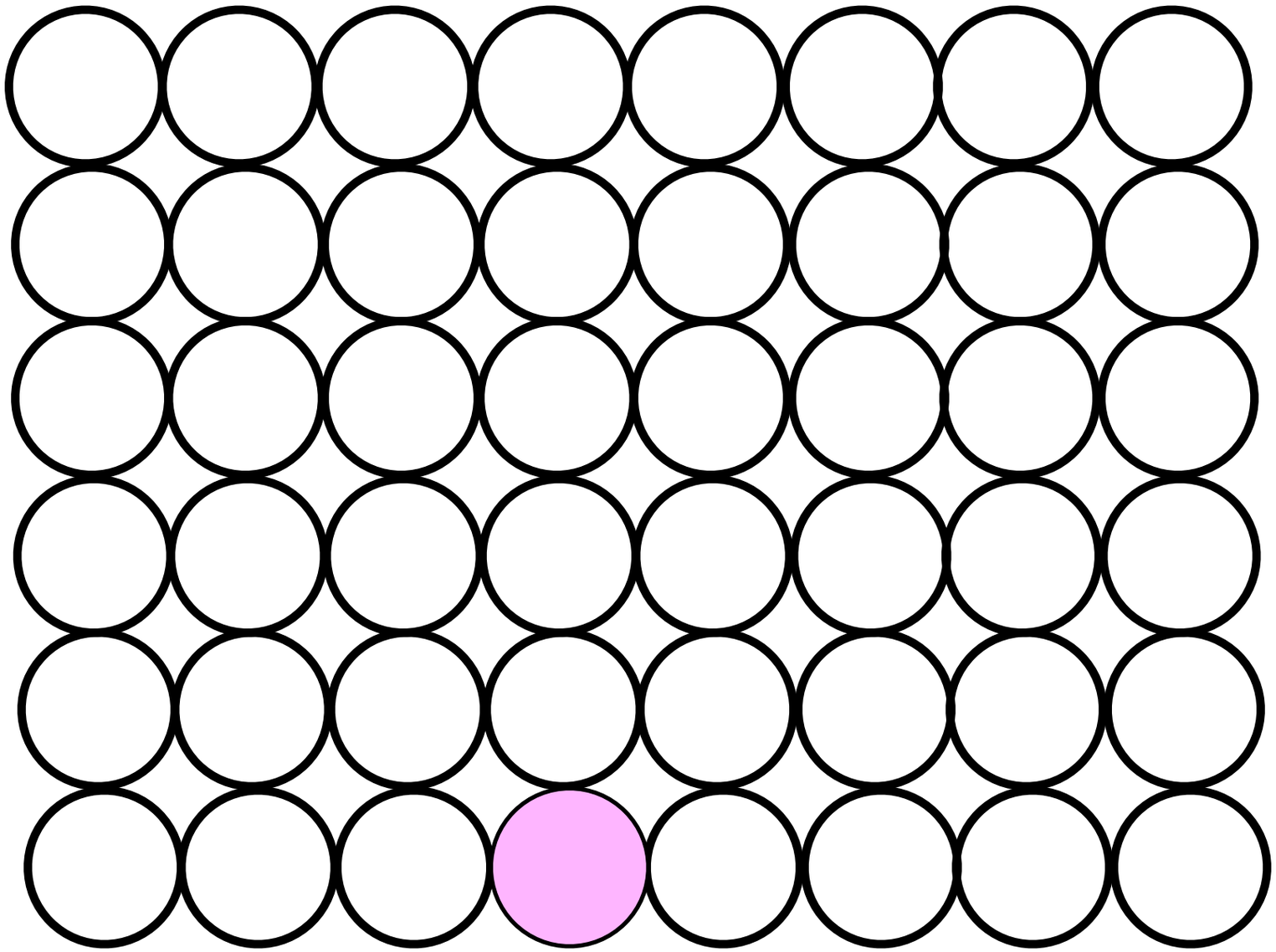
Past



Present

Coalescent model within 1 population

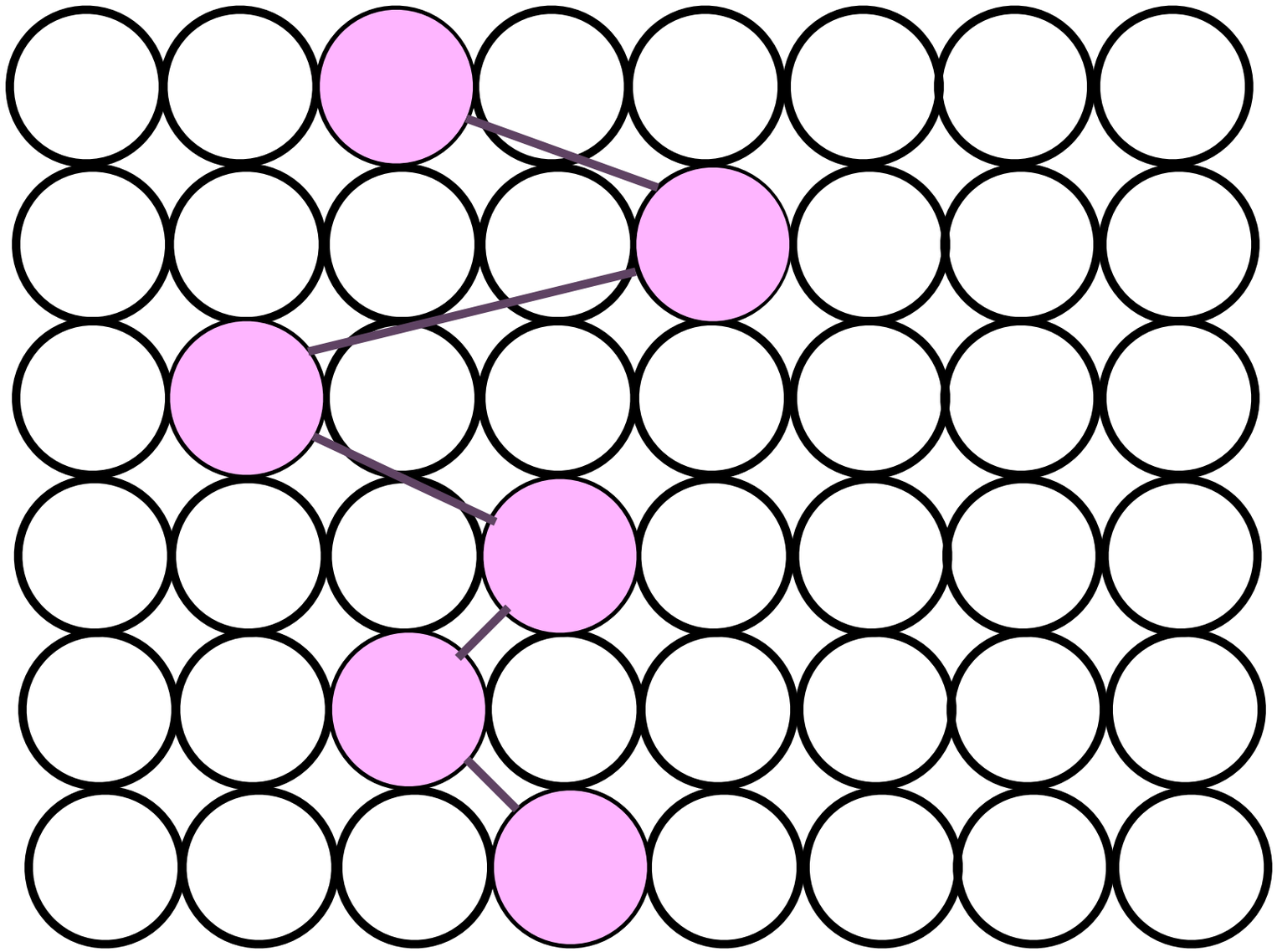
Past



Present

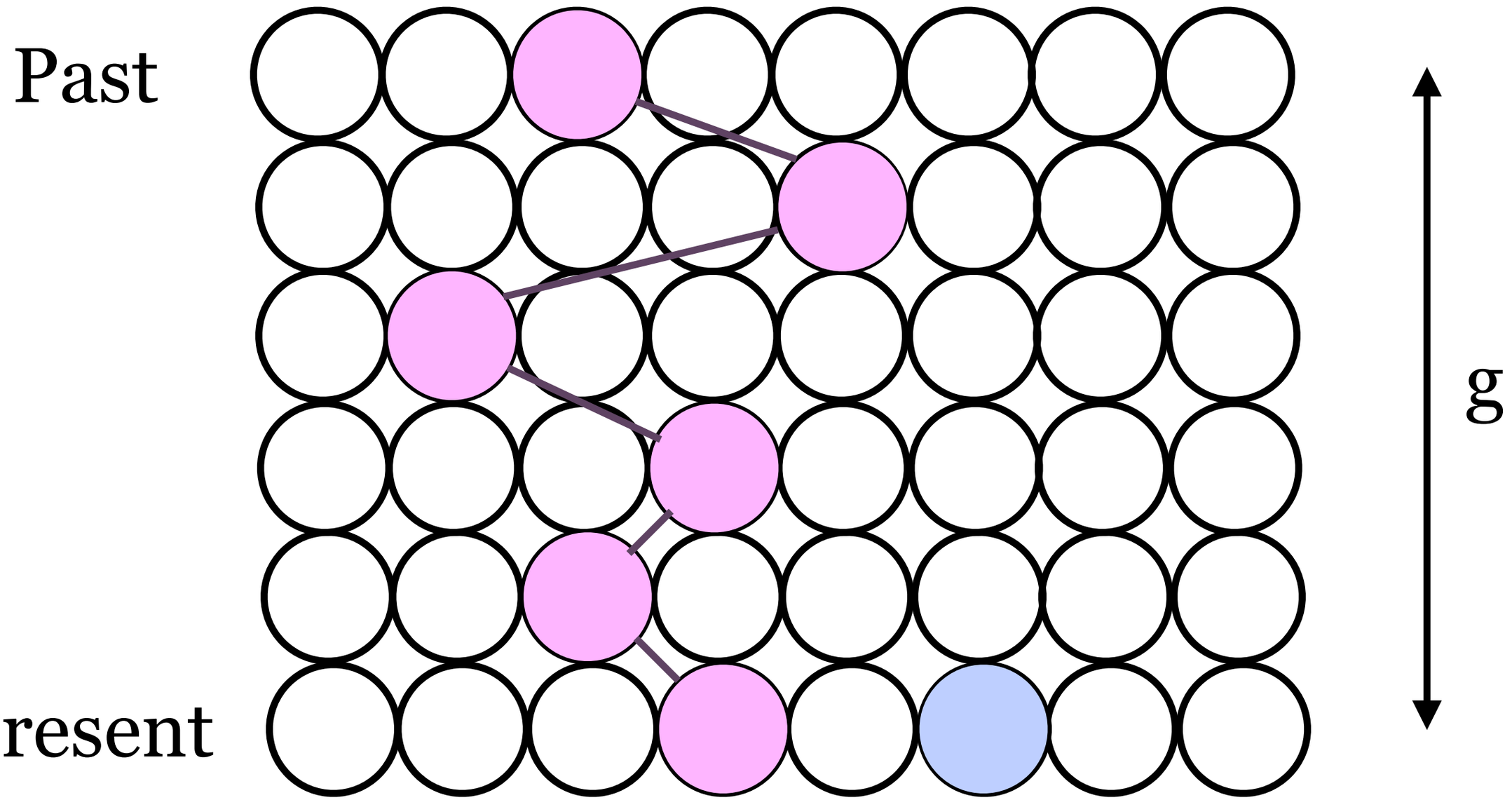
Coalescent model within 1 population

Past

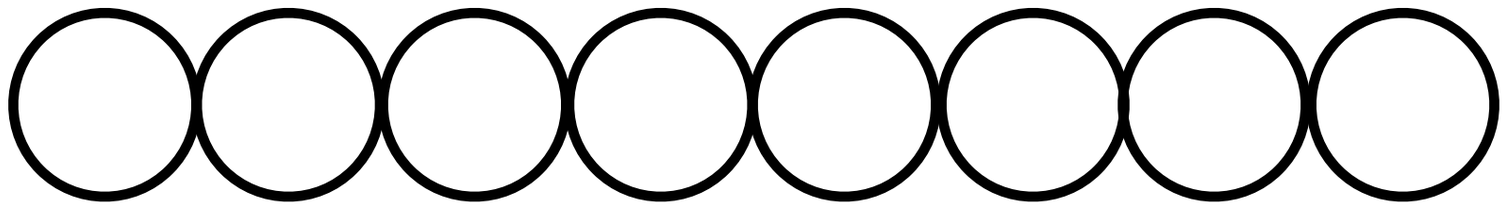


Present

Coalescent model within 1 population



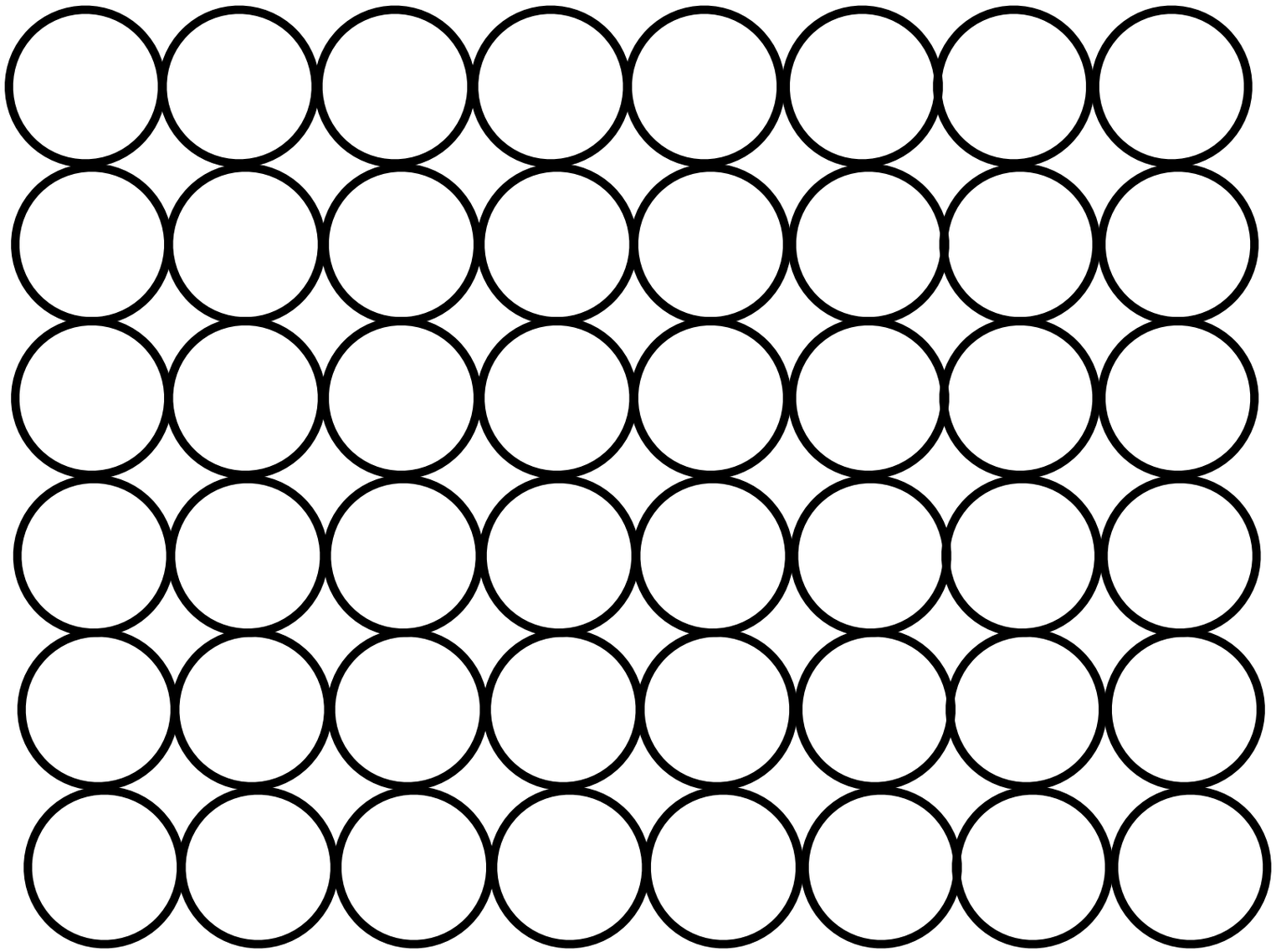
Coalescent model within 1 population

Present 

How many generations do we have to wait for these two individuals to reach a common ancestor?

Coalescent model within 1 population

Past

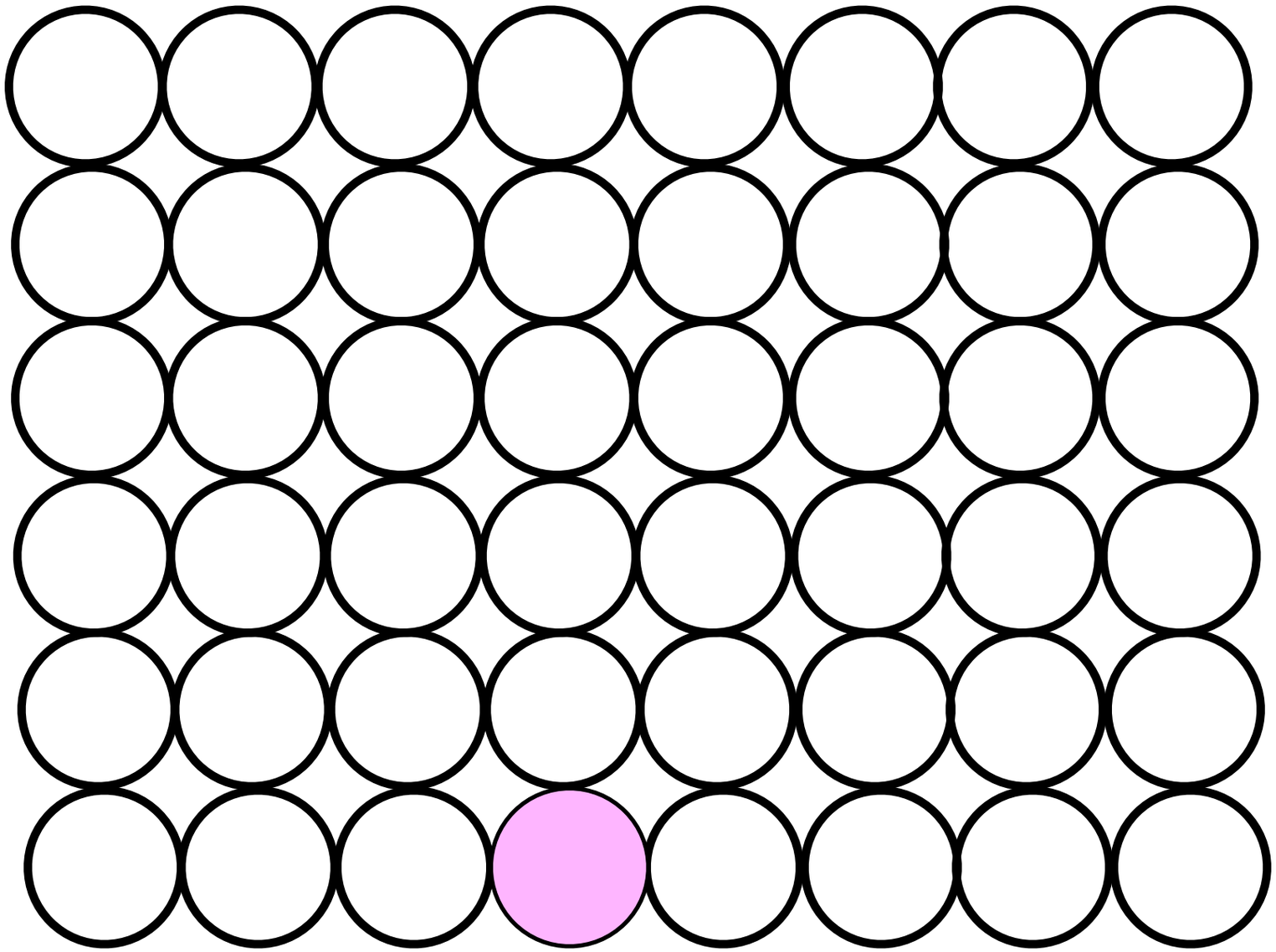


Present

How many generations do we have to wait for these two individuals to reach a common ancestor?

Coalescent model within 1 population

Past

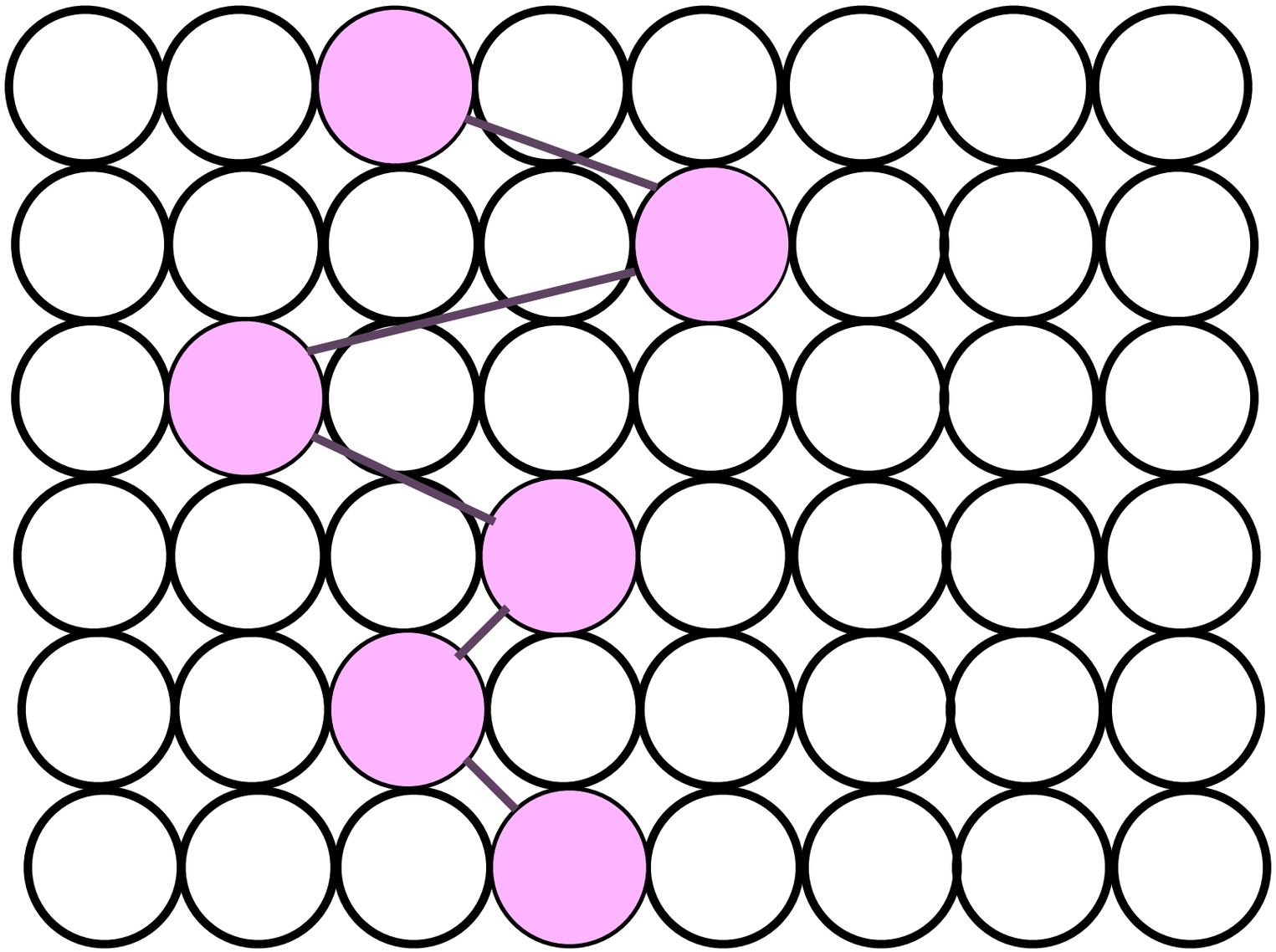


Present

How many generations do we have to wait for these two individuals to reach a common ancestor?

Coalescent model within 1 population

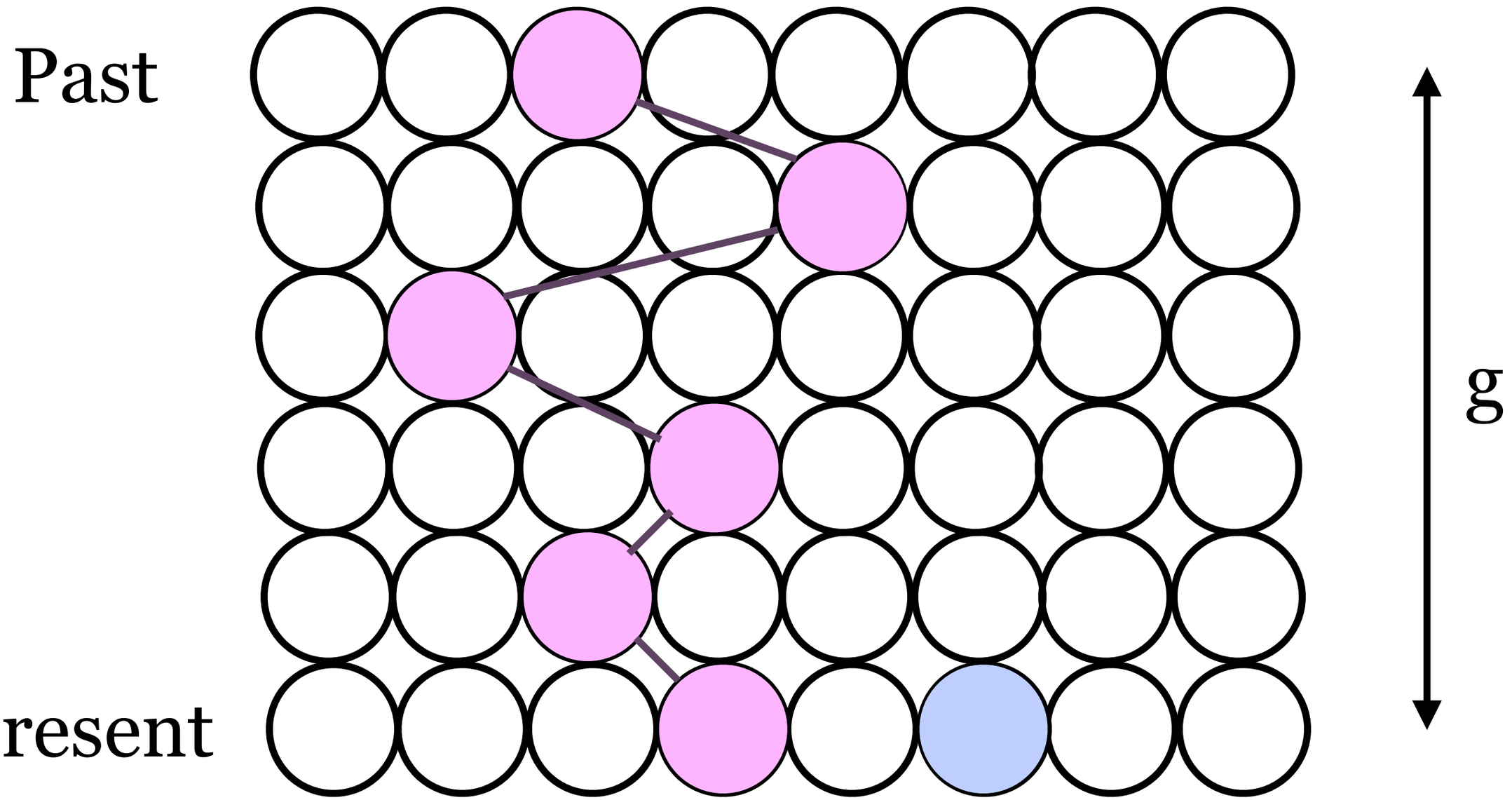
Past



Present

How many generations do we have to wait for these two individuals to reach a common ancestor?

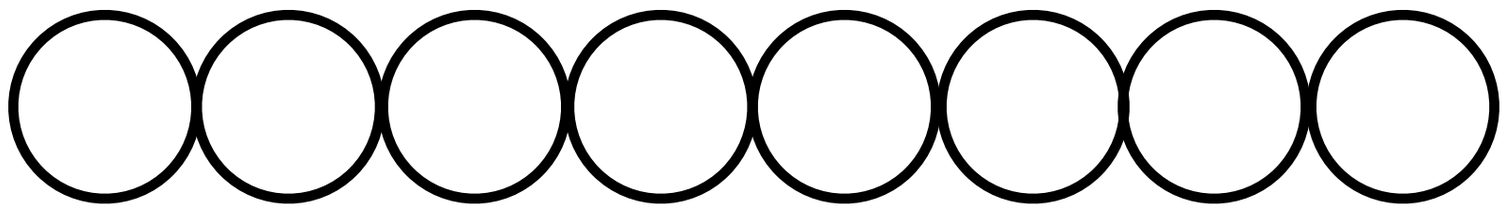
Coalescent model within 1 population



How many generations do we have to wait for these two individuals to reach a common ancestor?

Coalescent model within 1 population

Present



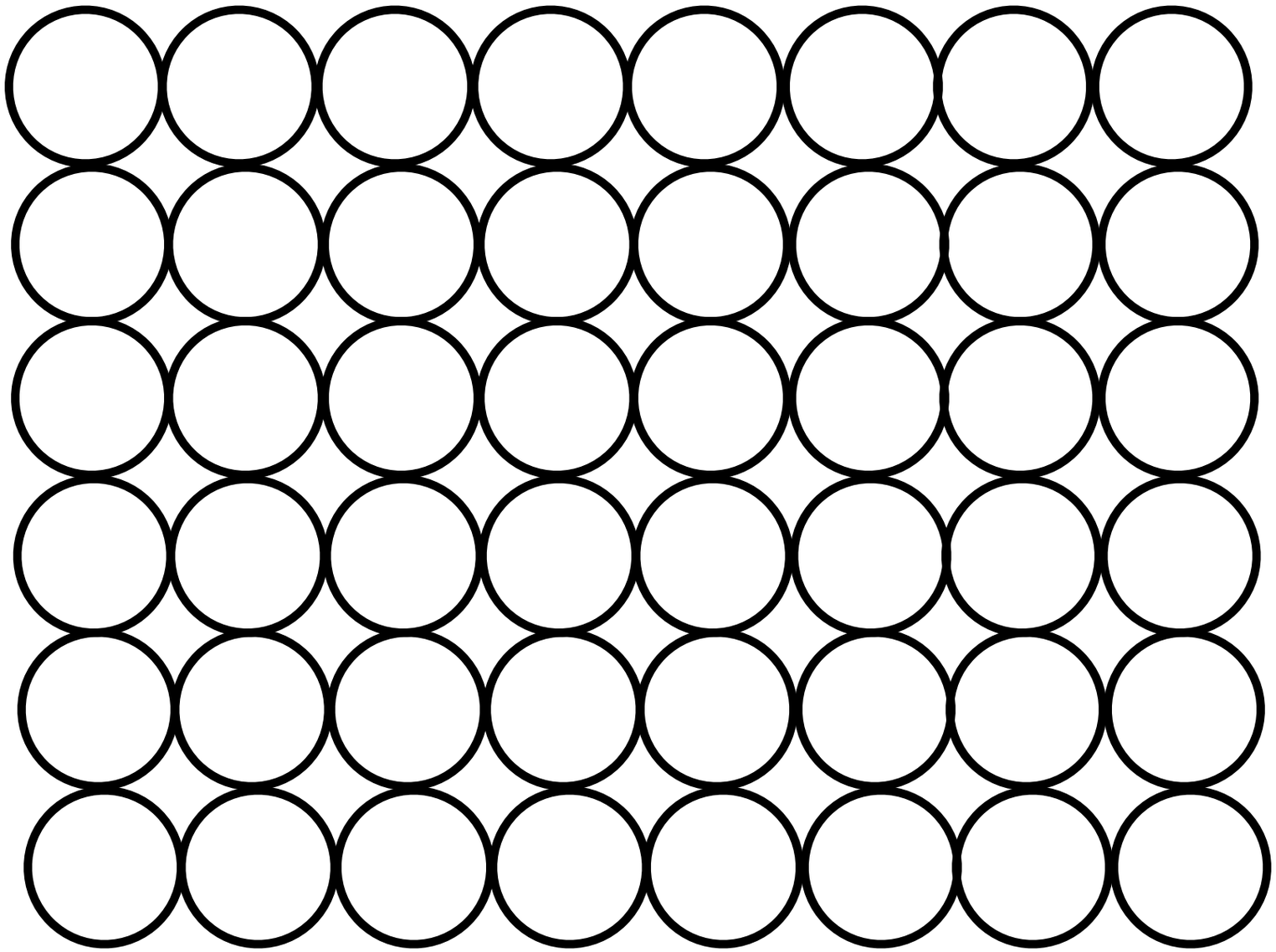
How many generations do we have to wait for these two individuals to reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

Coalescent model within 1 population

Past



Present

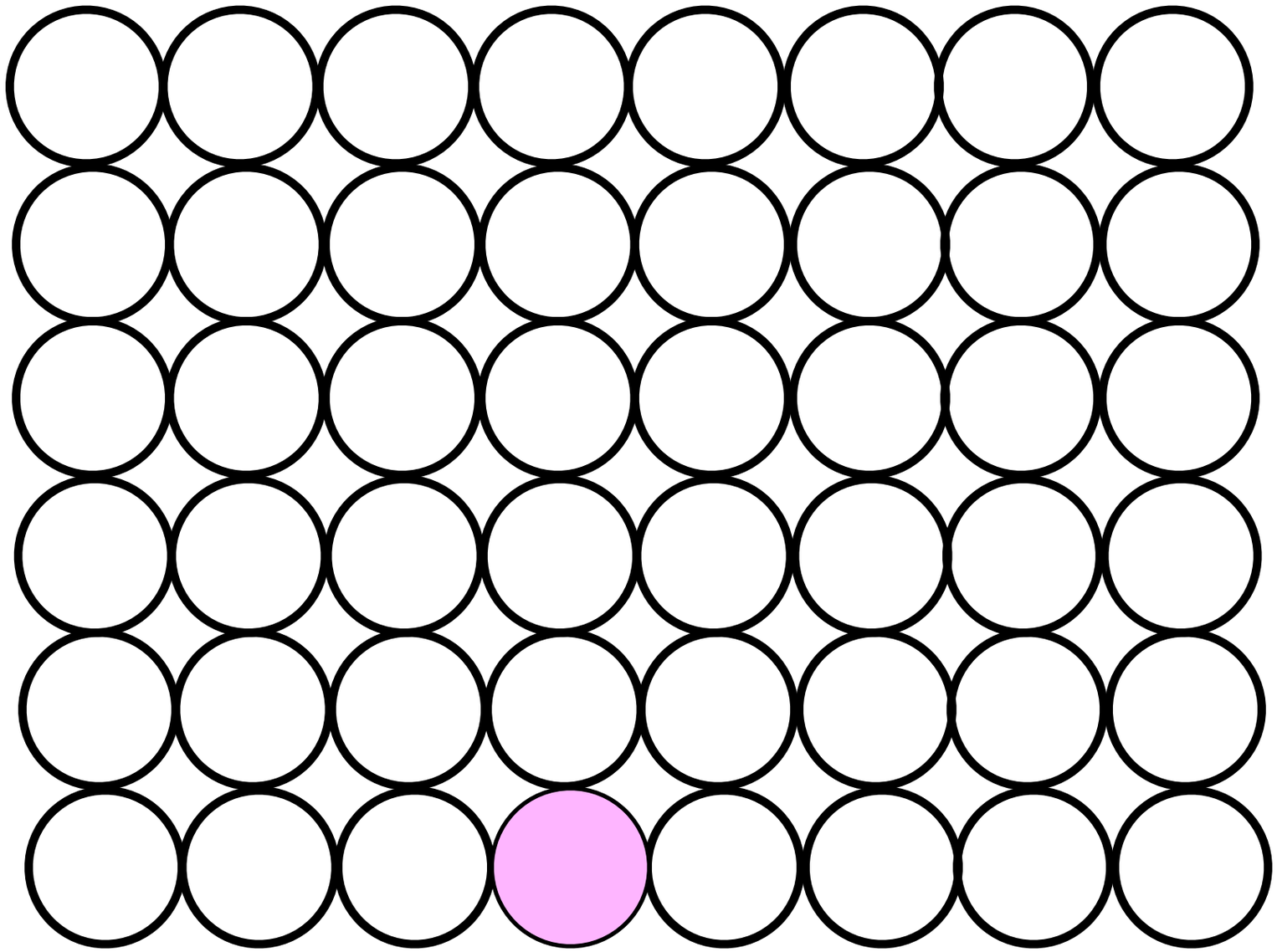
How many generations do we have to wait for these two individuals to reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

Coalescent model within 1 population

Past



Present

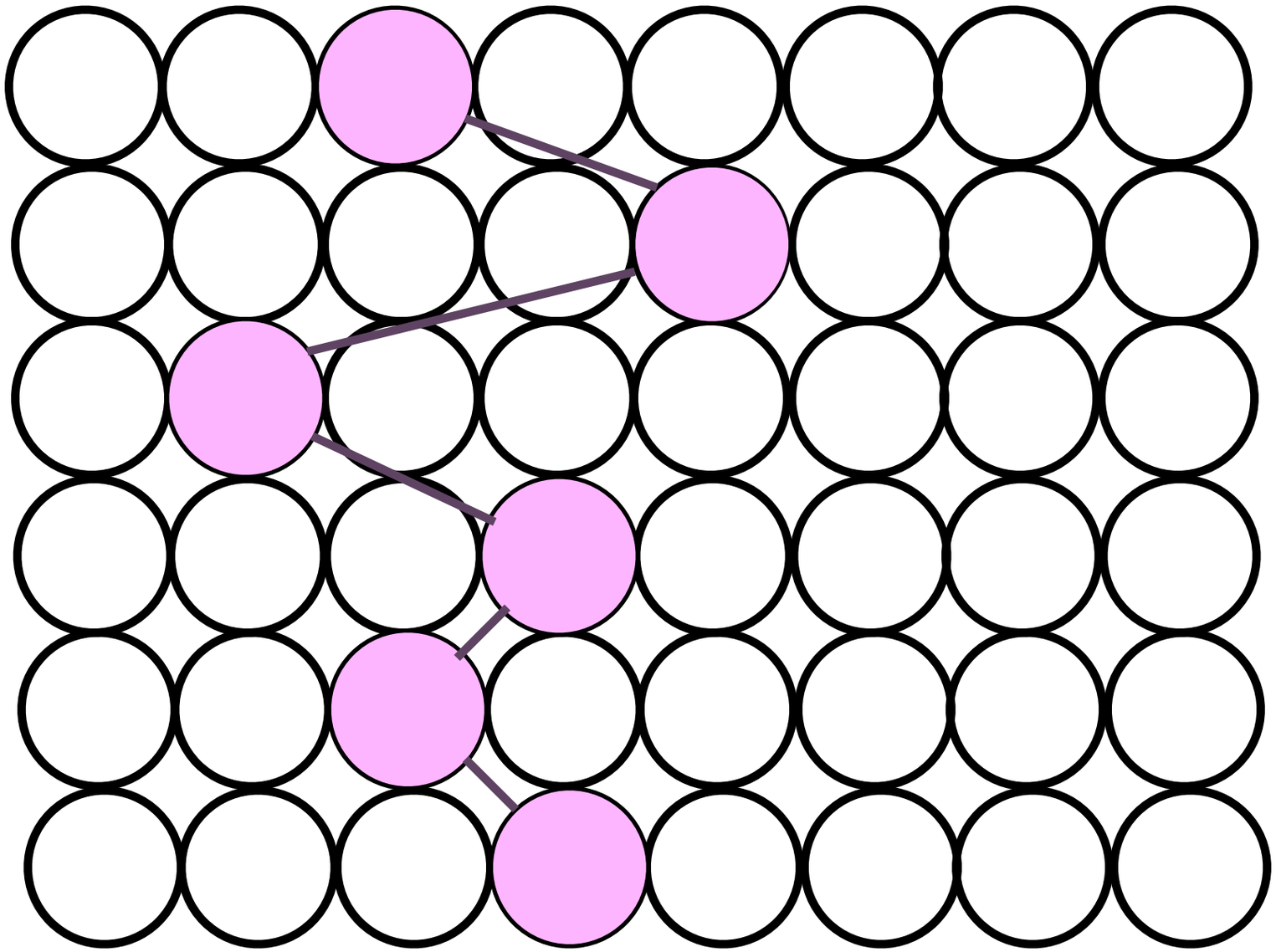
How many generations do we have to wait for these two individuals to reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

Coalescent model within 1 population

Past



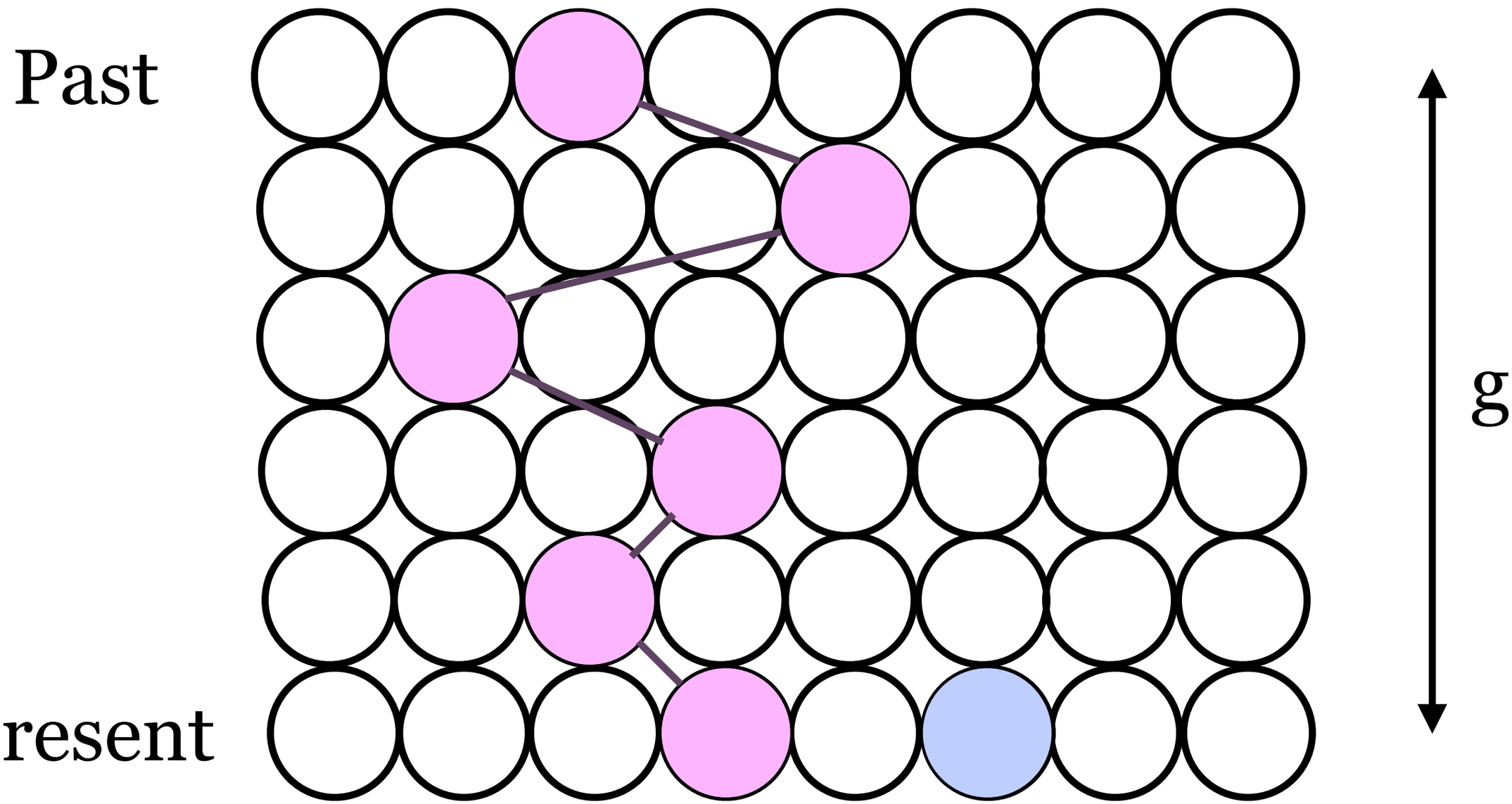
Present

How many generations do we have to wait for these two individuals to reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

Coalescent model within 1 population



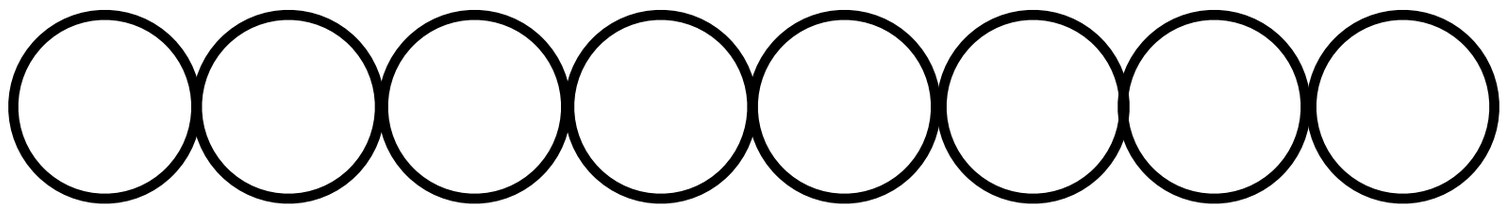
How many generations do we have to wait for these two individuals to reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

Coalescent model within 1 population

Present



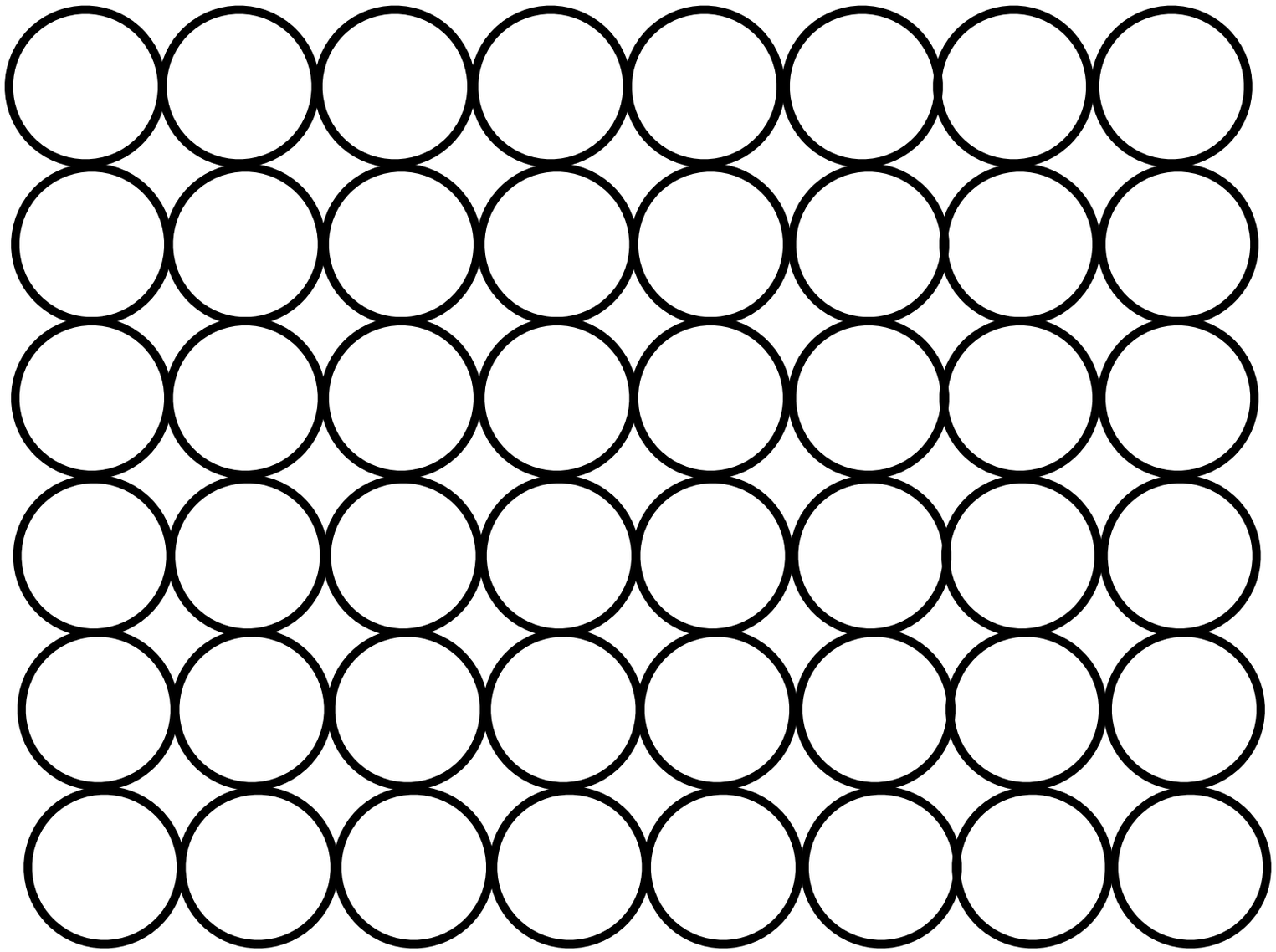
How many generations do we have to wait for these two individuals to reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

Coalescent model within 1 population

Past



Present

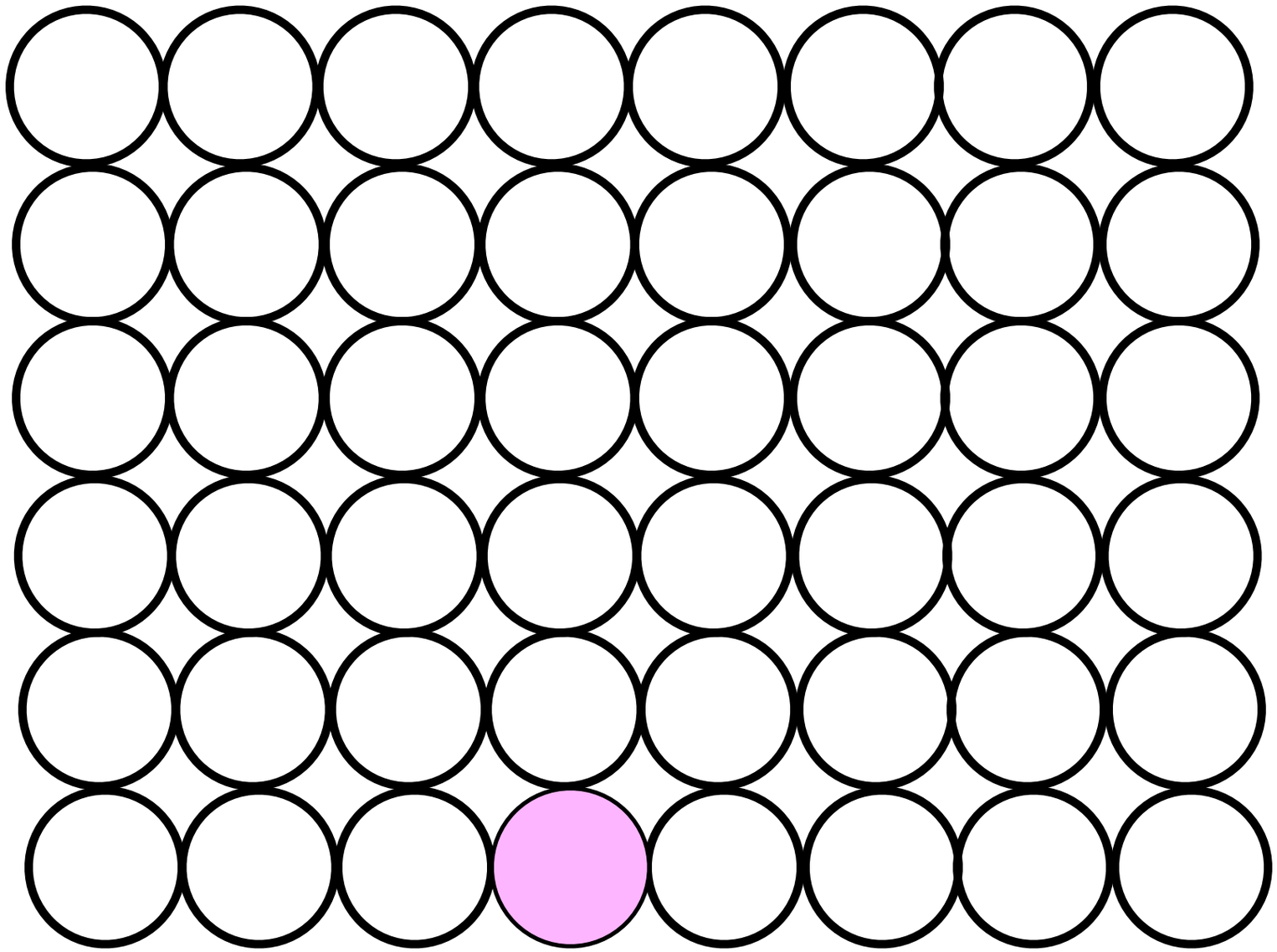
How many generations do we have to wait for these two individuals to reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

Coalescent model within 1 population

Past



Present

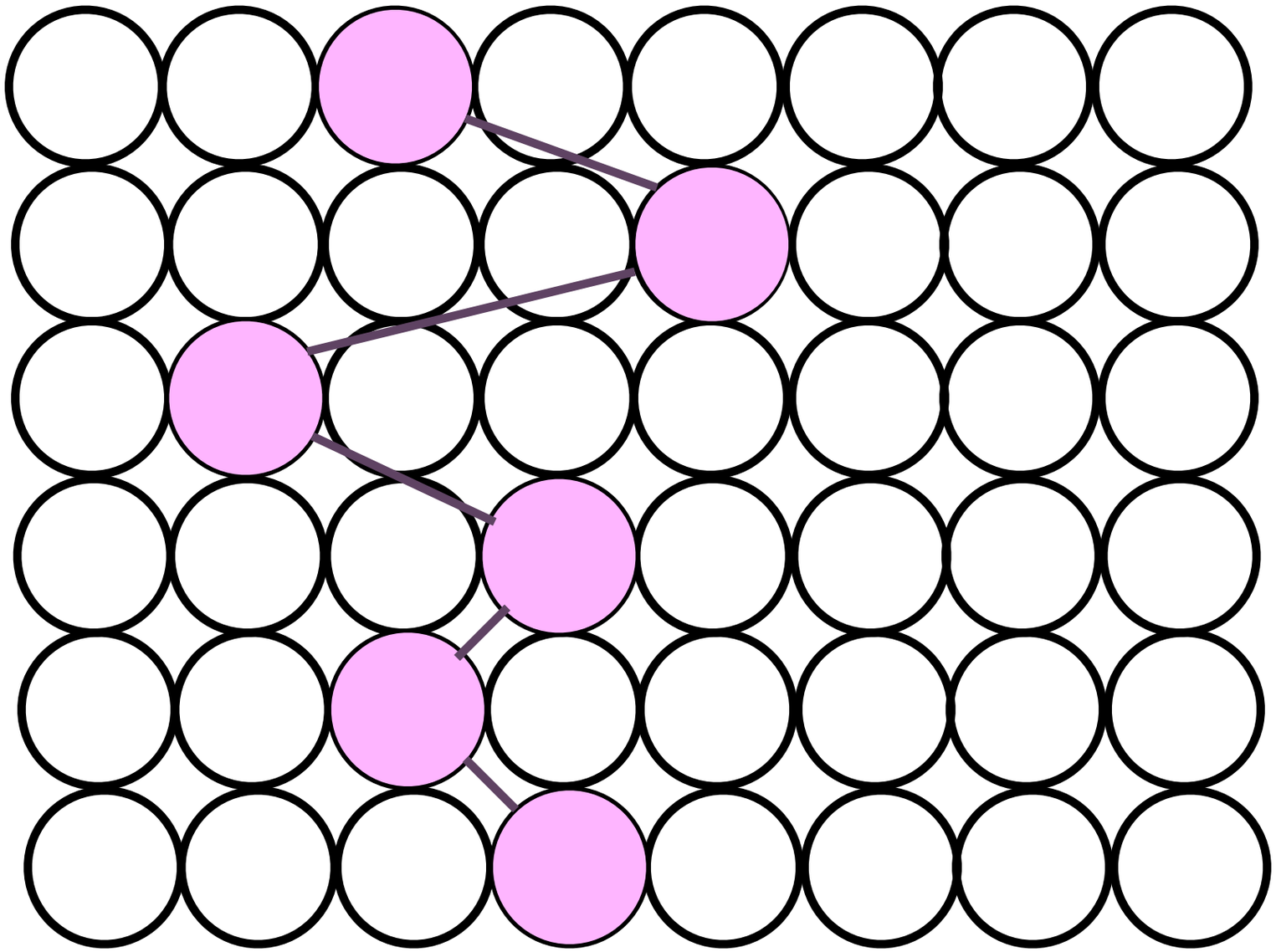
How many generations do we have to wait for these two individuals to reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

Coalescent model within 1 population

Past



Present

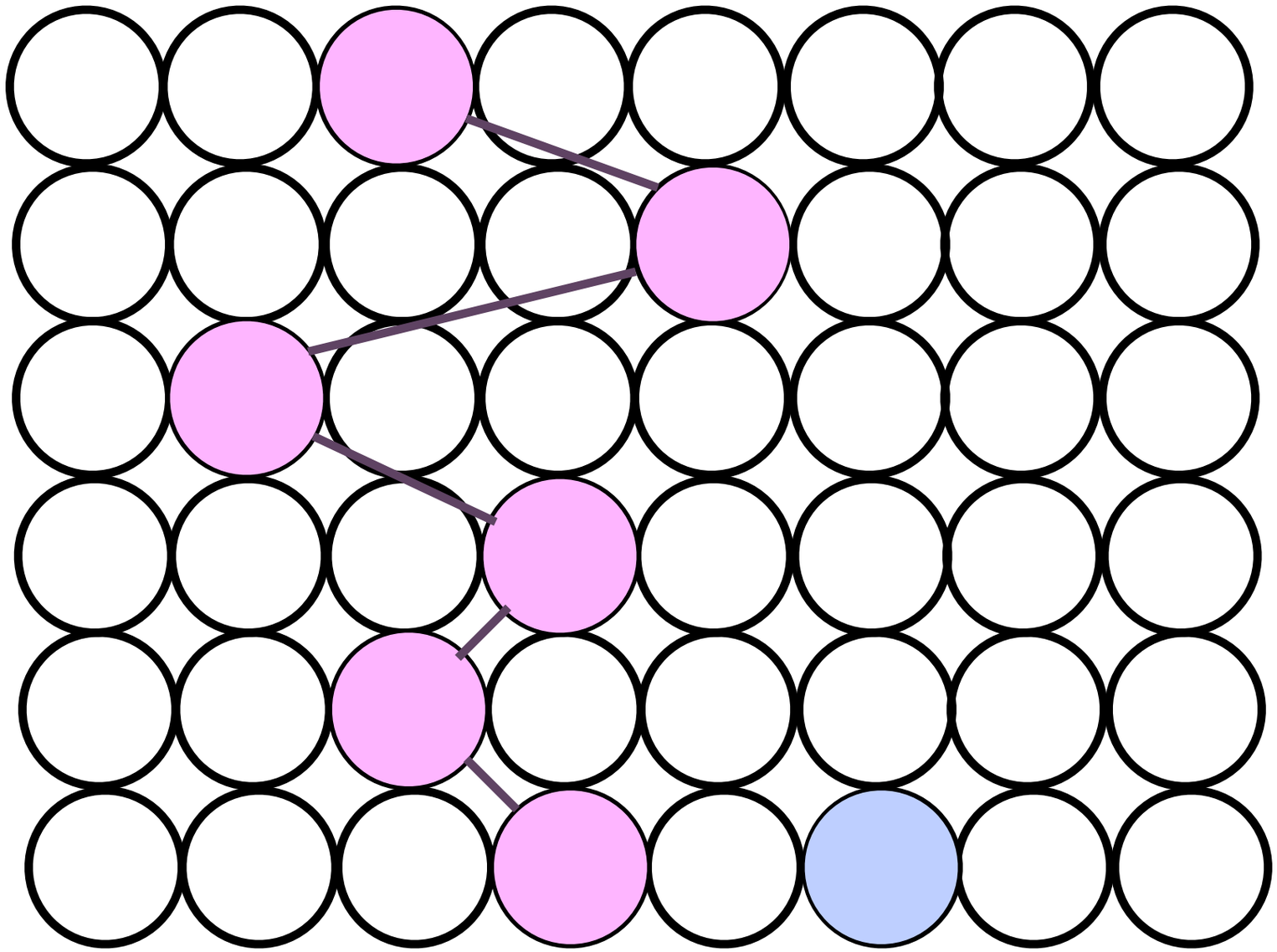
How many generations do we have to wait for these two individuals to reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

Coalescent model within 1 population

Past



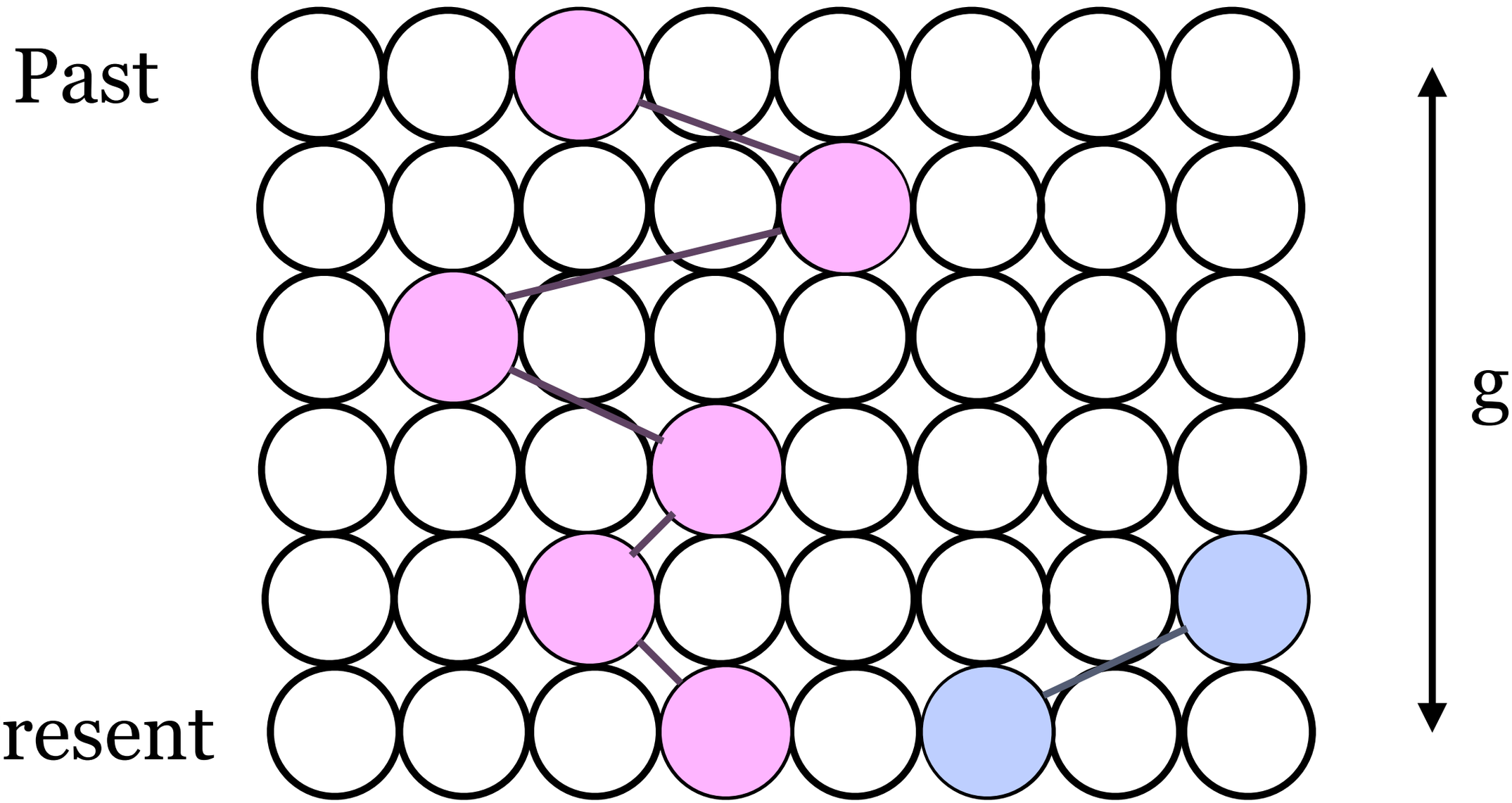
Present

How many generations do we have to wait for these two individuals to reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

Coalescent model within 1 population

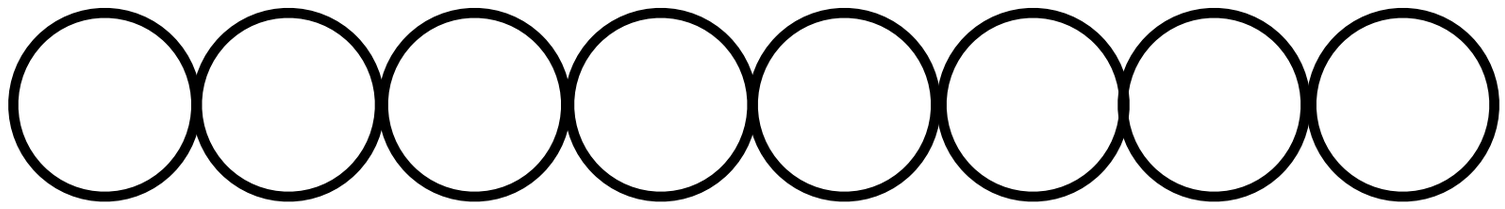


How many generations do we have to wait for these two individuals to reach a common ancestor?

$$P(\text{coalesce}) = \frac{1}{N}$$

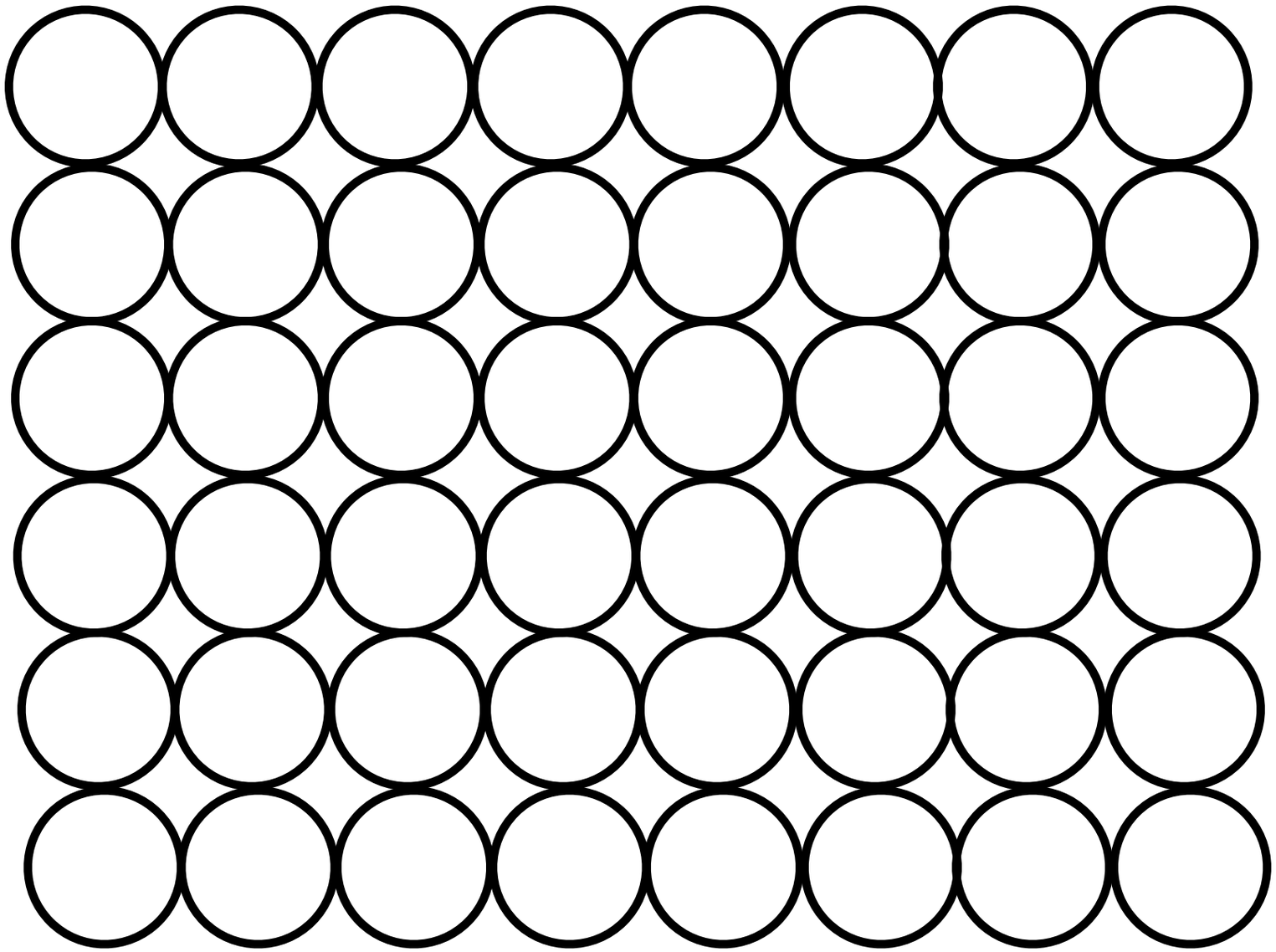
$$P(\text{no coalesce}) = 1 - \frac{1}{N}$$

Coalescent model within 1 population

Present 

Coalescent model within 1 population

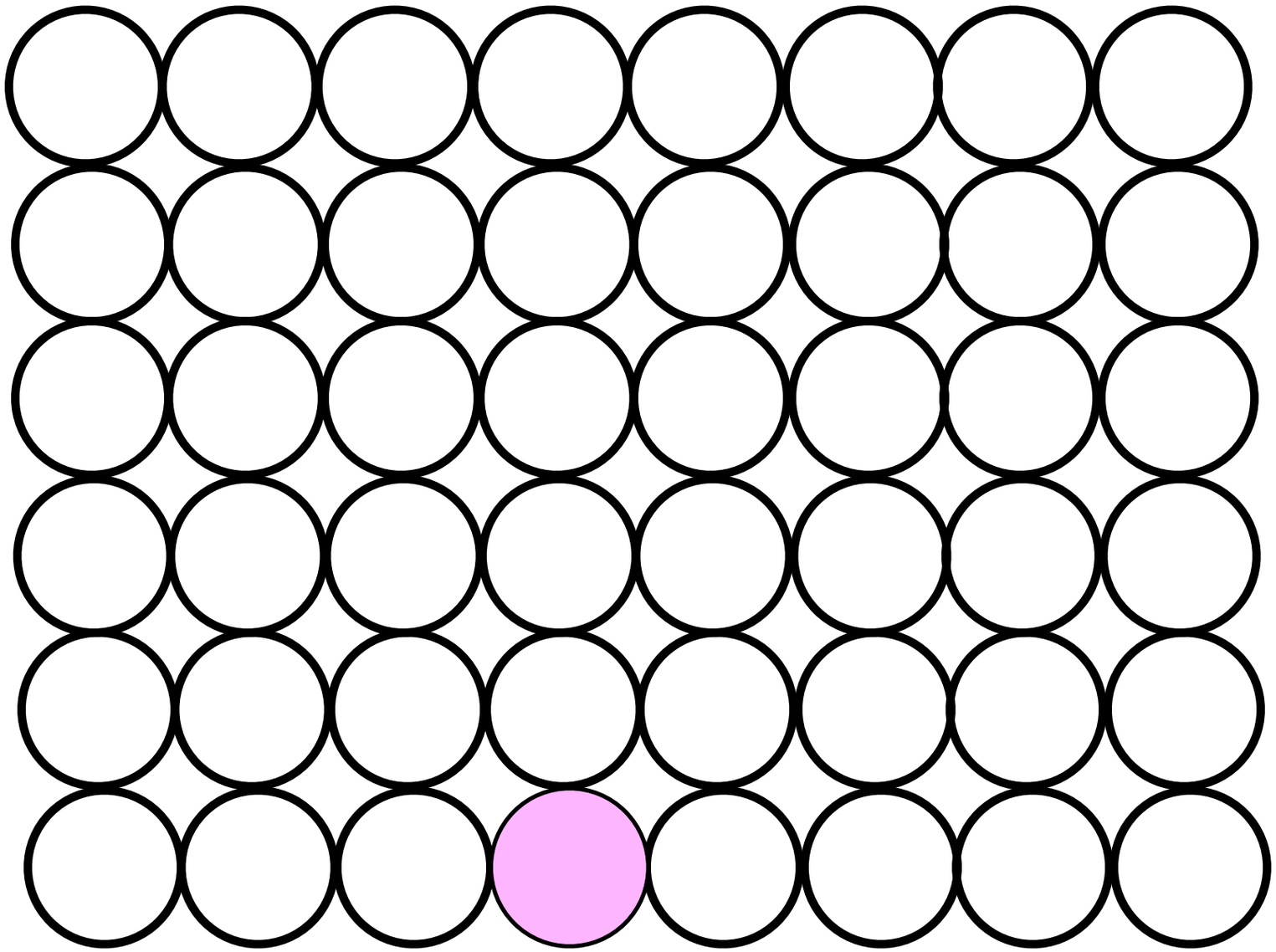
Past



Present

Coalescent model within 1 population

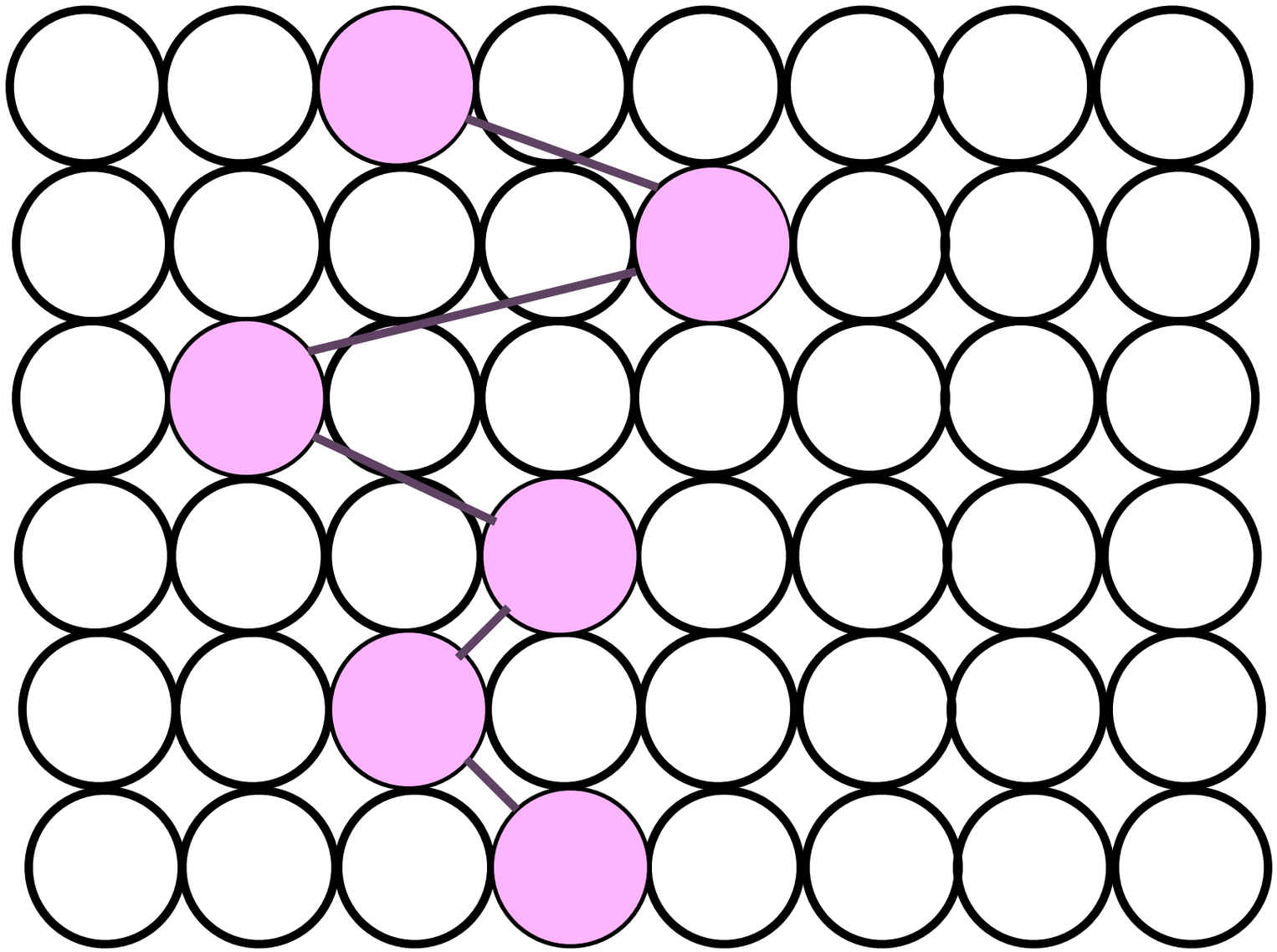
Past



Present

Coalescent model within 1 population

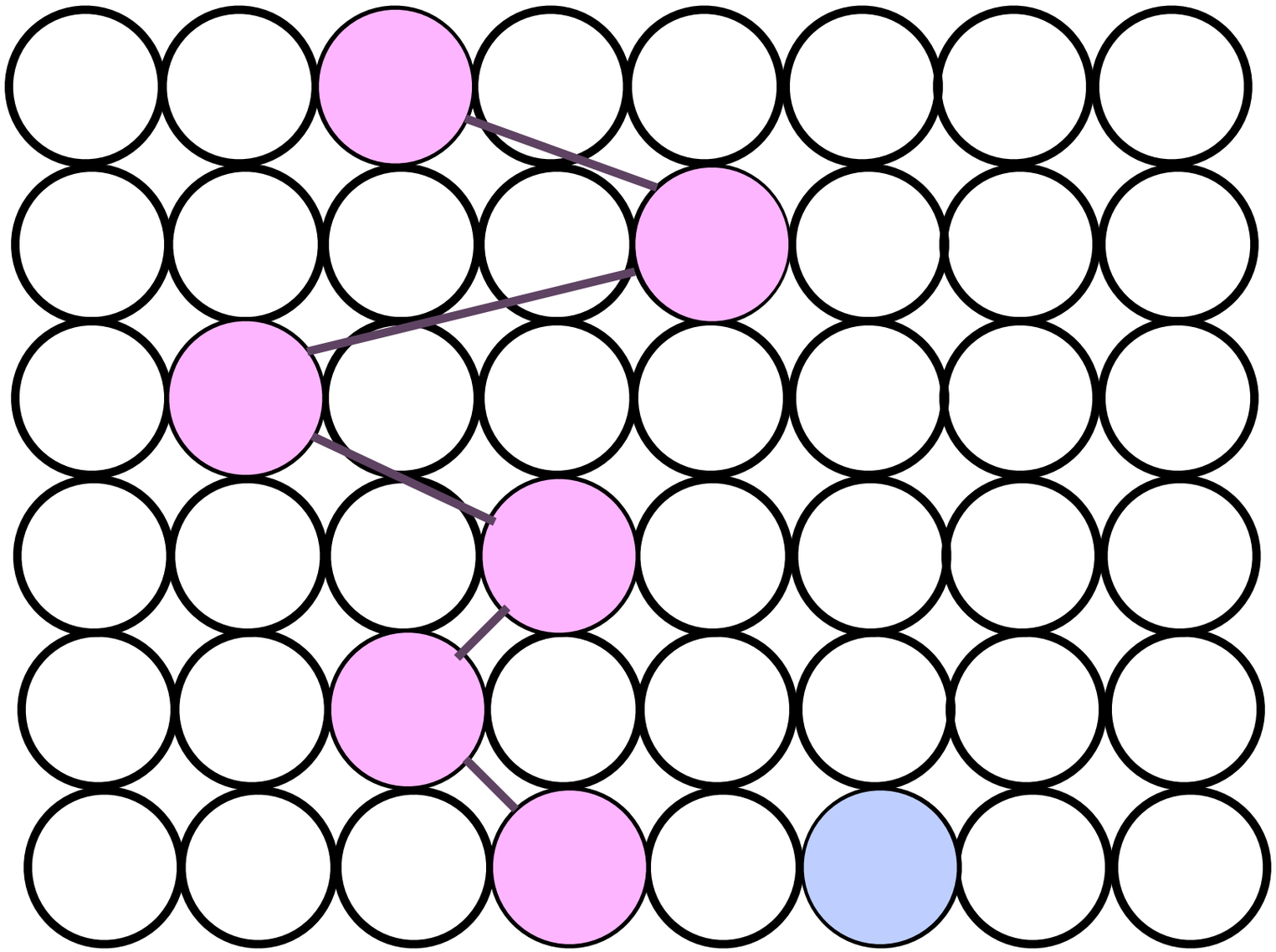
Past



Present

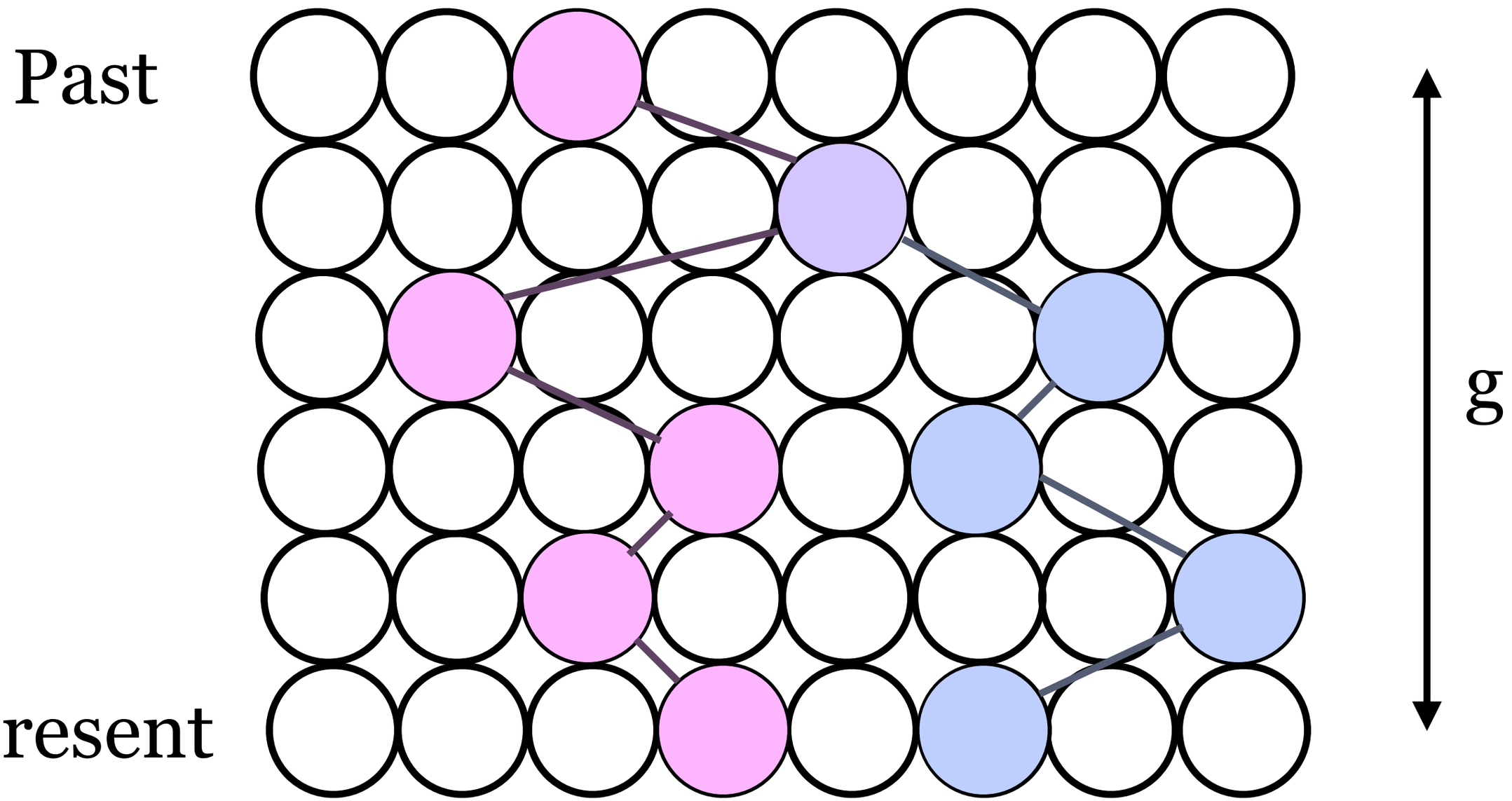
Coalescent model within 1 population

Past

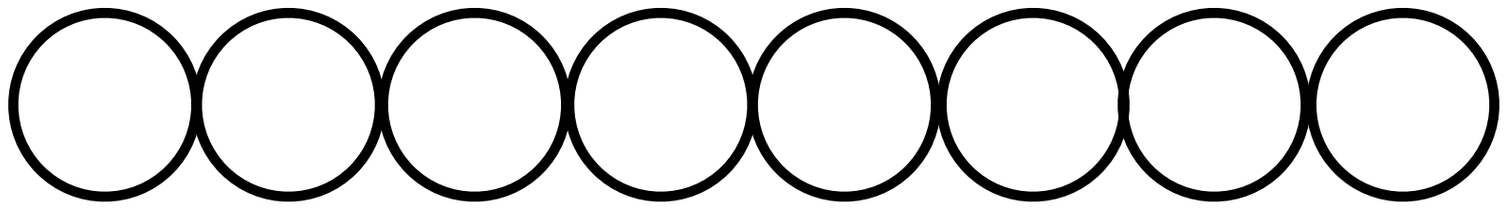


Present

Coalescent model within 1 population

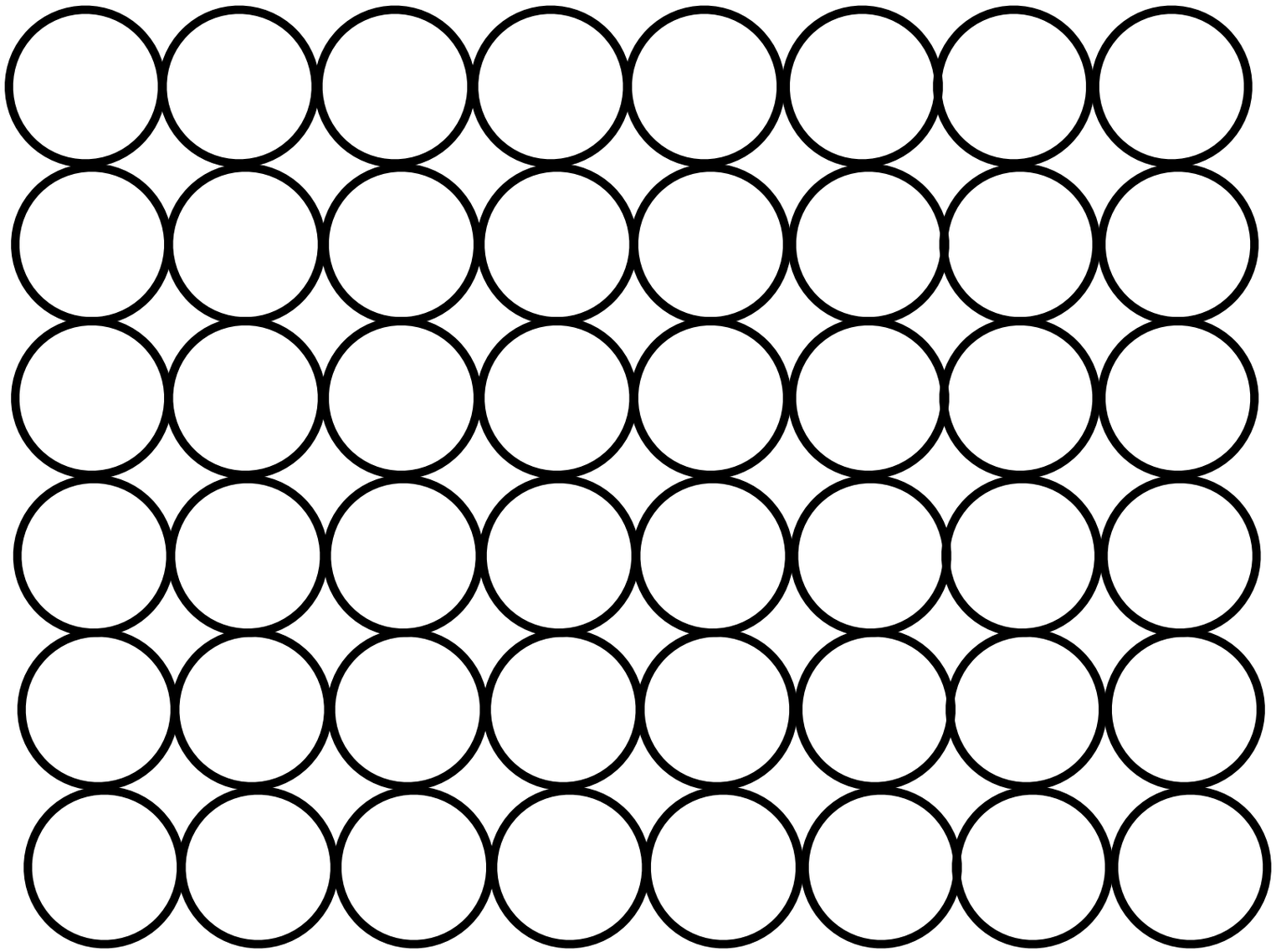


Coalescent model within 1 population

Present 

Coalescent model within 1 population

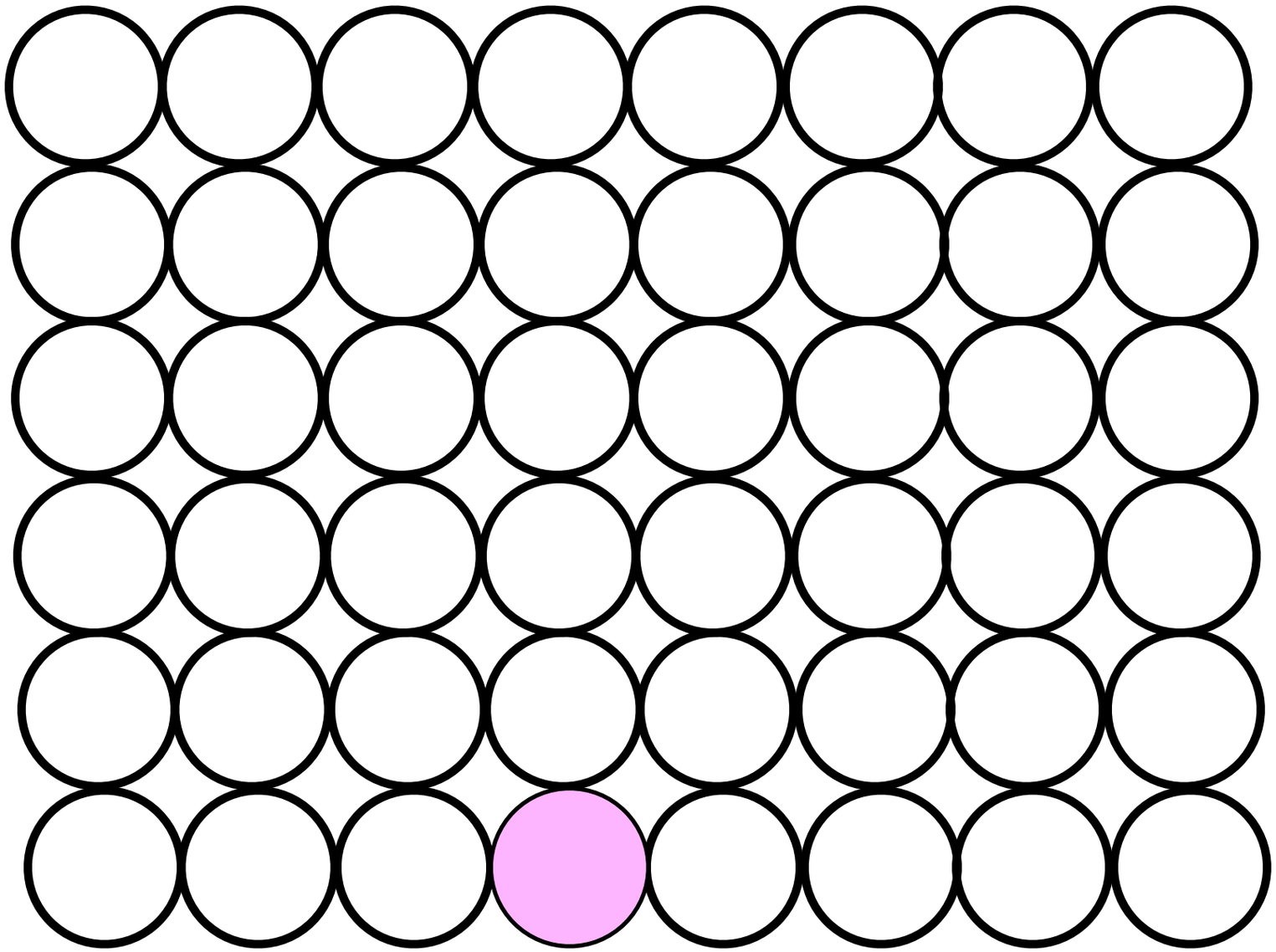
Past



Present

Coalescent model within 1 population

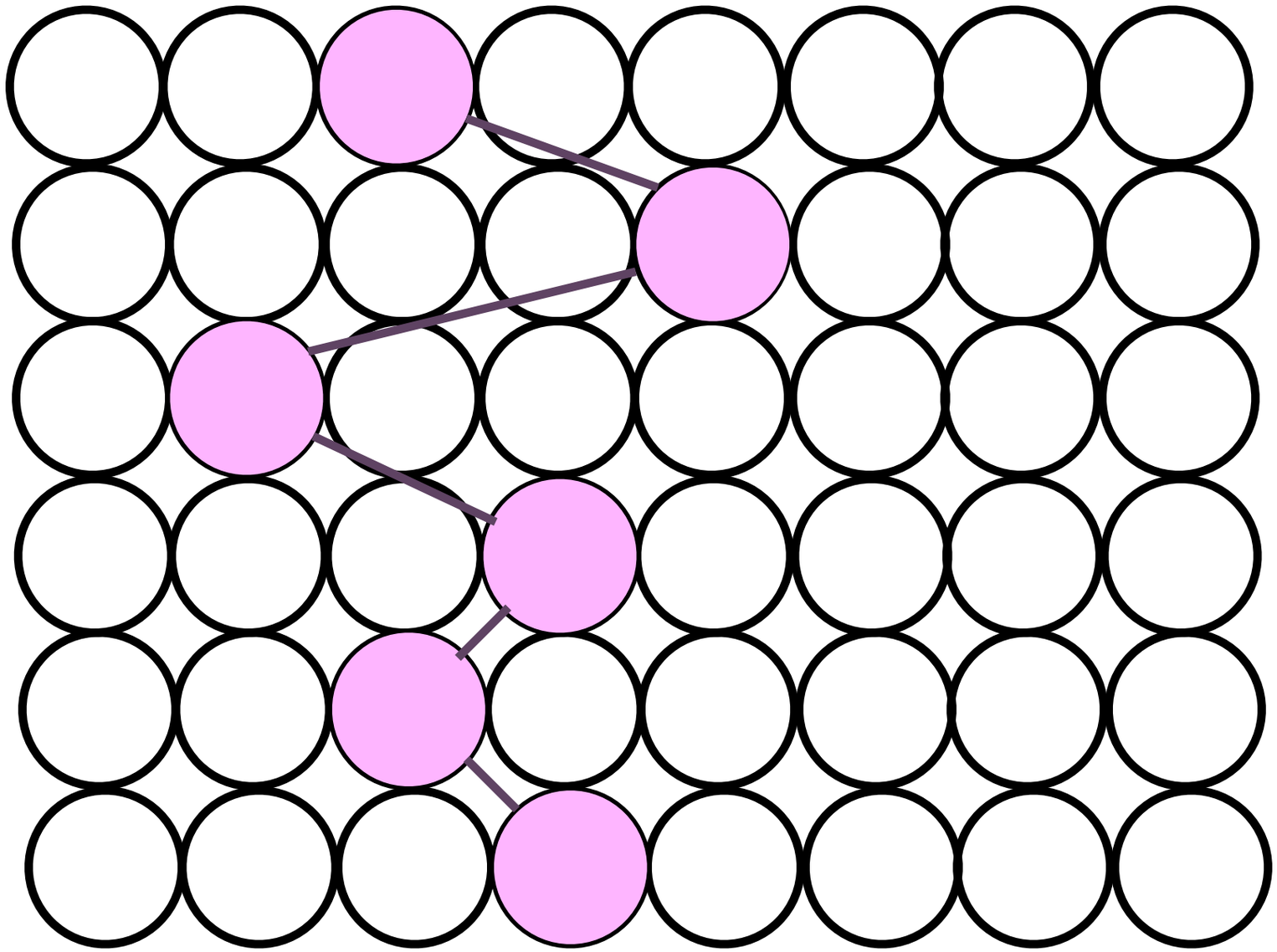
Past



Present

Coalescent model within 1 population

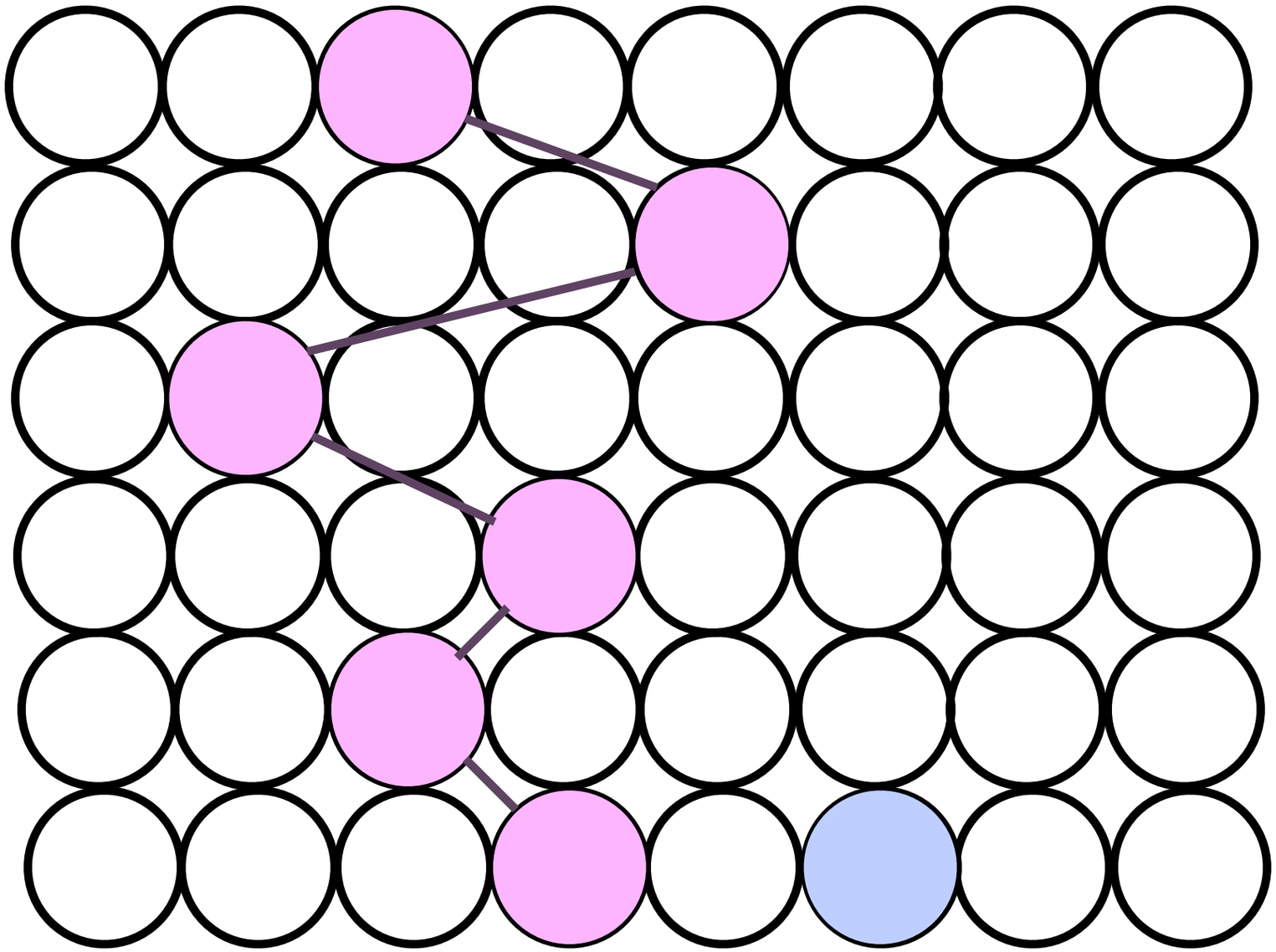
Past



Present

Coalescent model within 1 population

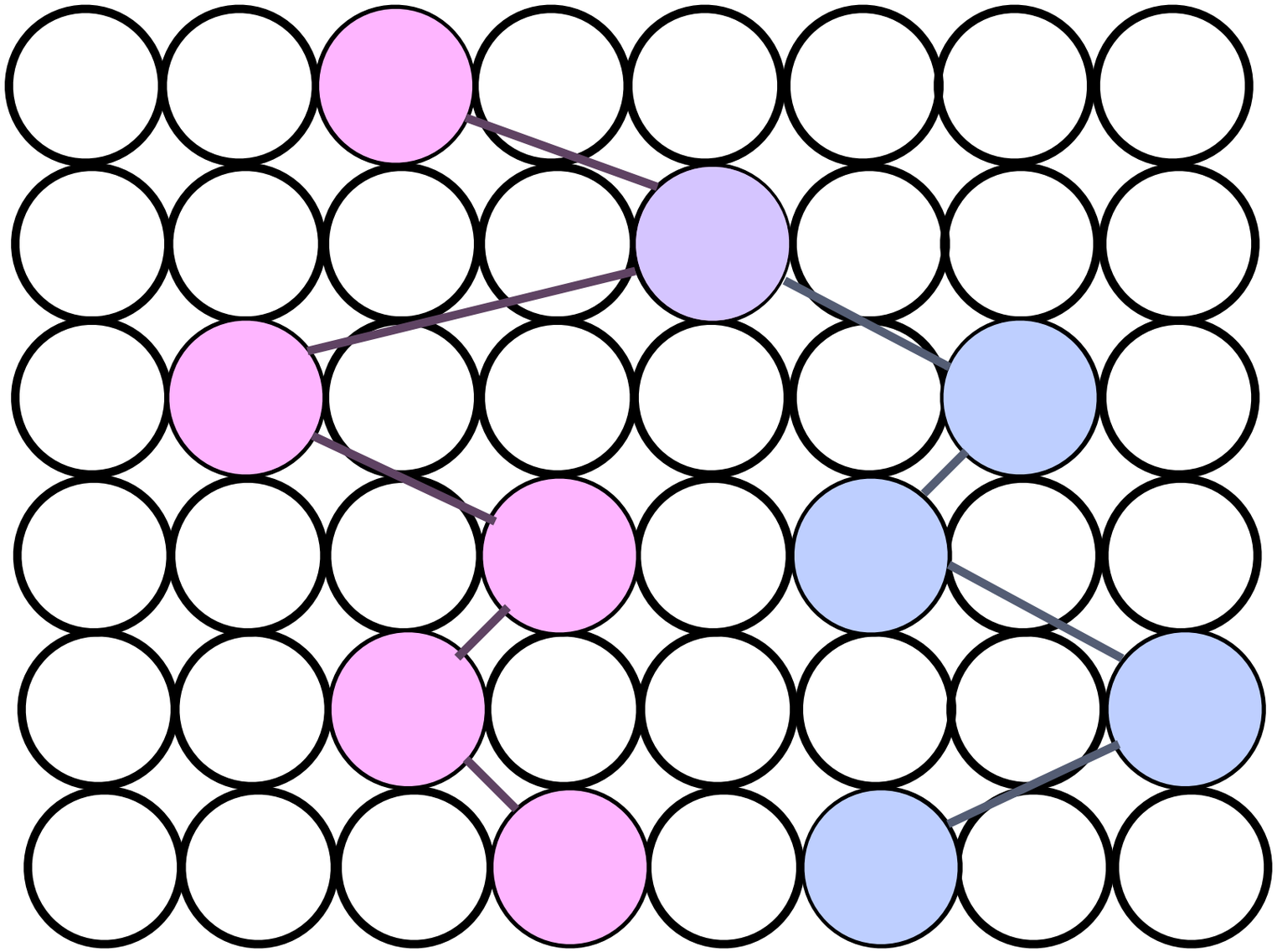
Past



Present

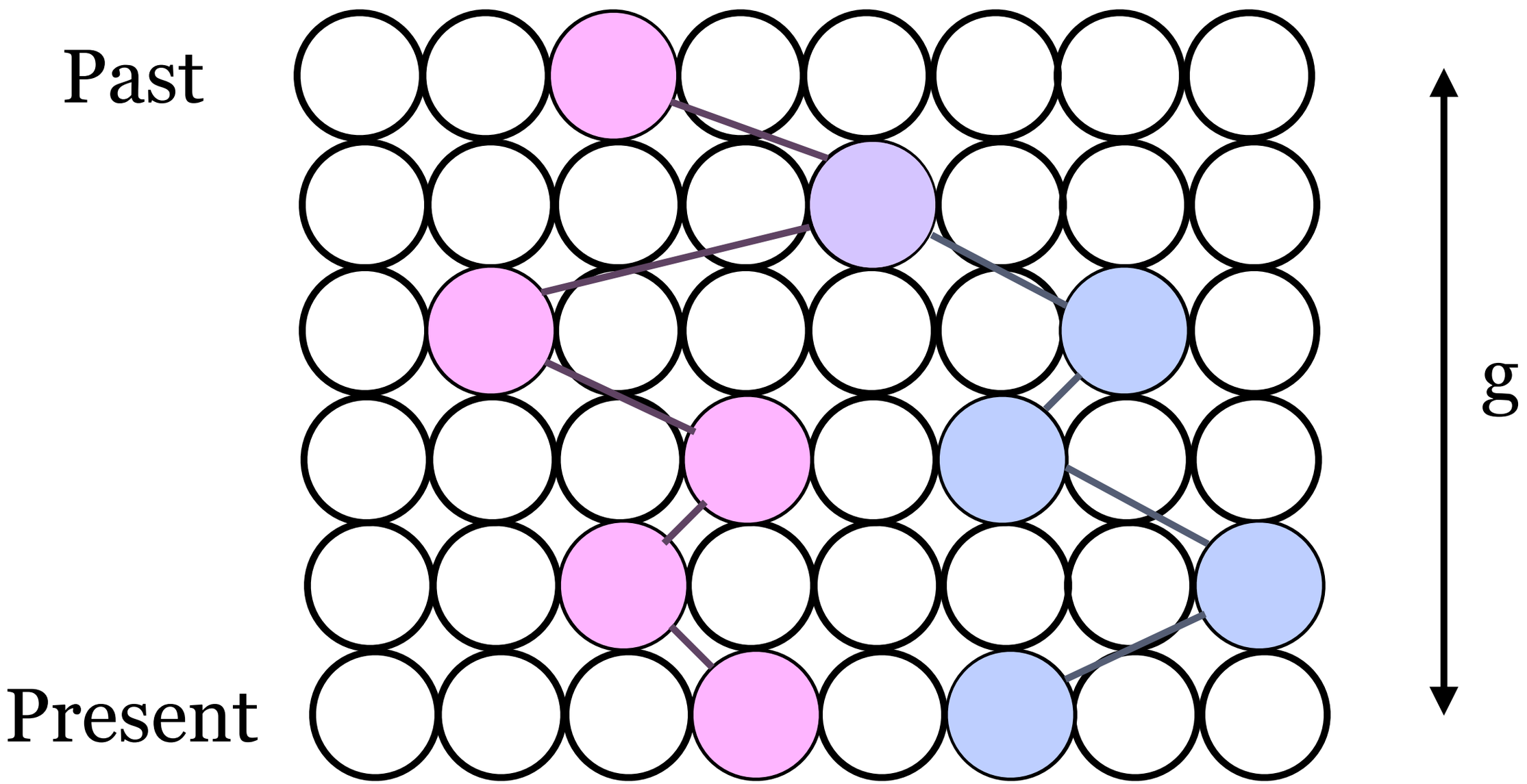
Coalescent model within 1 population

Past



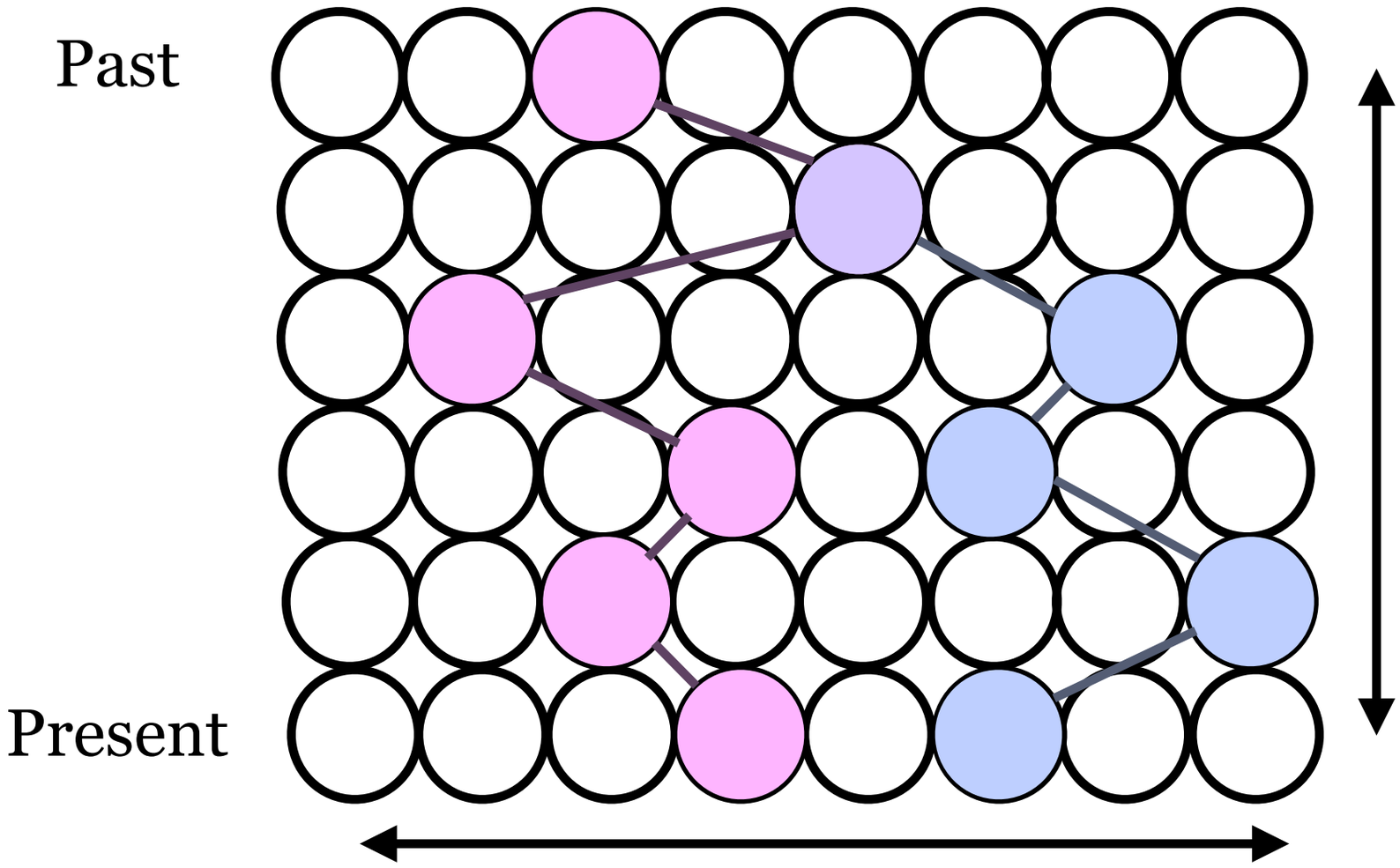
Present

Coalescent model within 1 population



Coalescent event

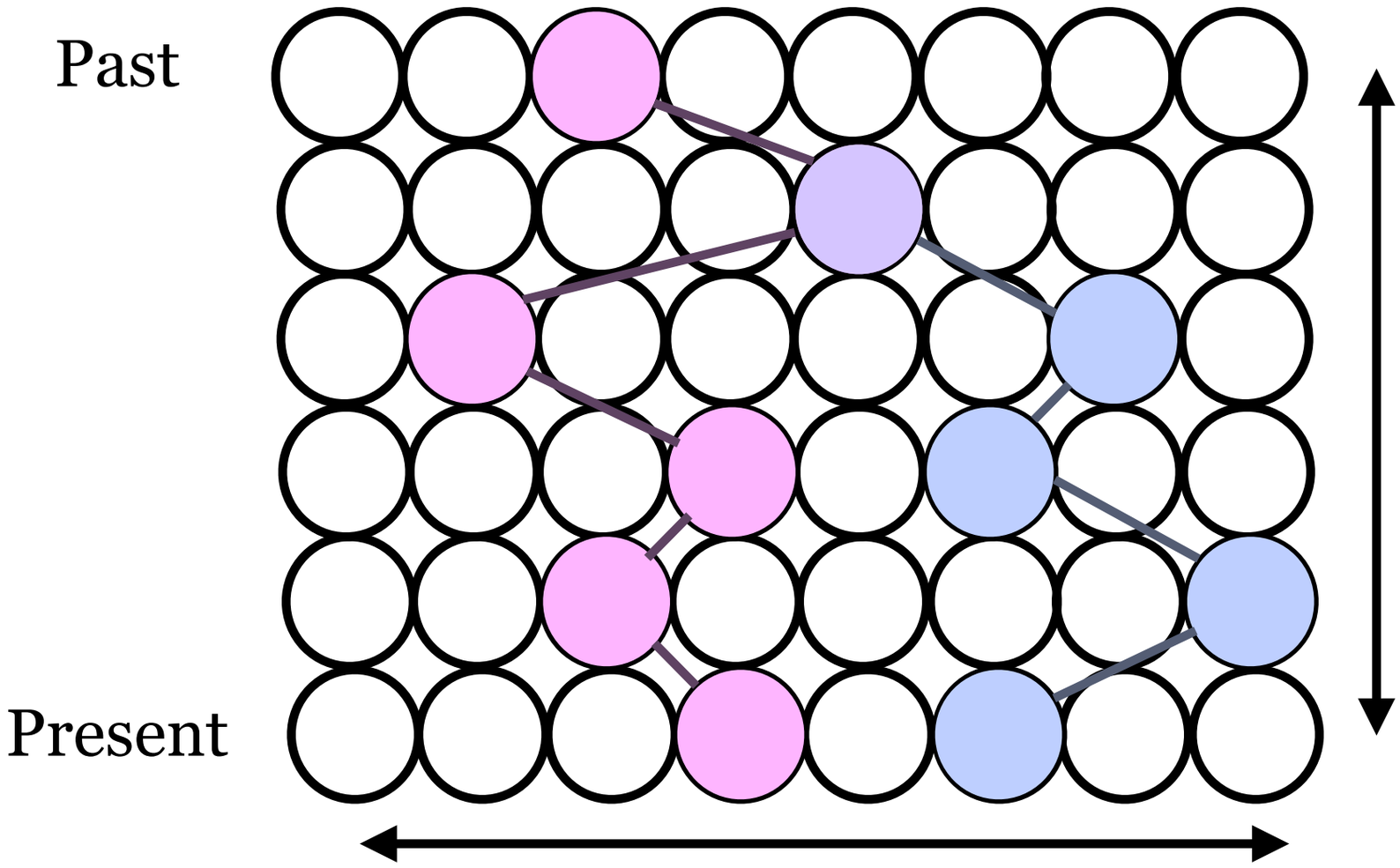
Coalescent model within 1 population



Probability of no coalescence in g generations: $\left(1-\frac{1}{N}\right)^{g}$

$$t = g/N \Rightarrow \left(1-\frac{t}{Nt}\right)^{Nt} \xrightarrow[N\to\infty]{} e^{-t}$$

Coalescent model within 1 population

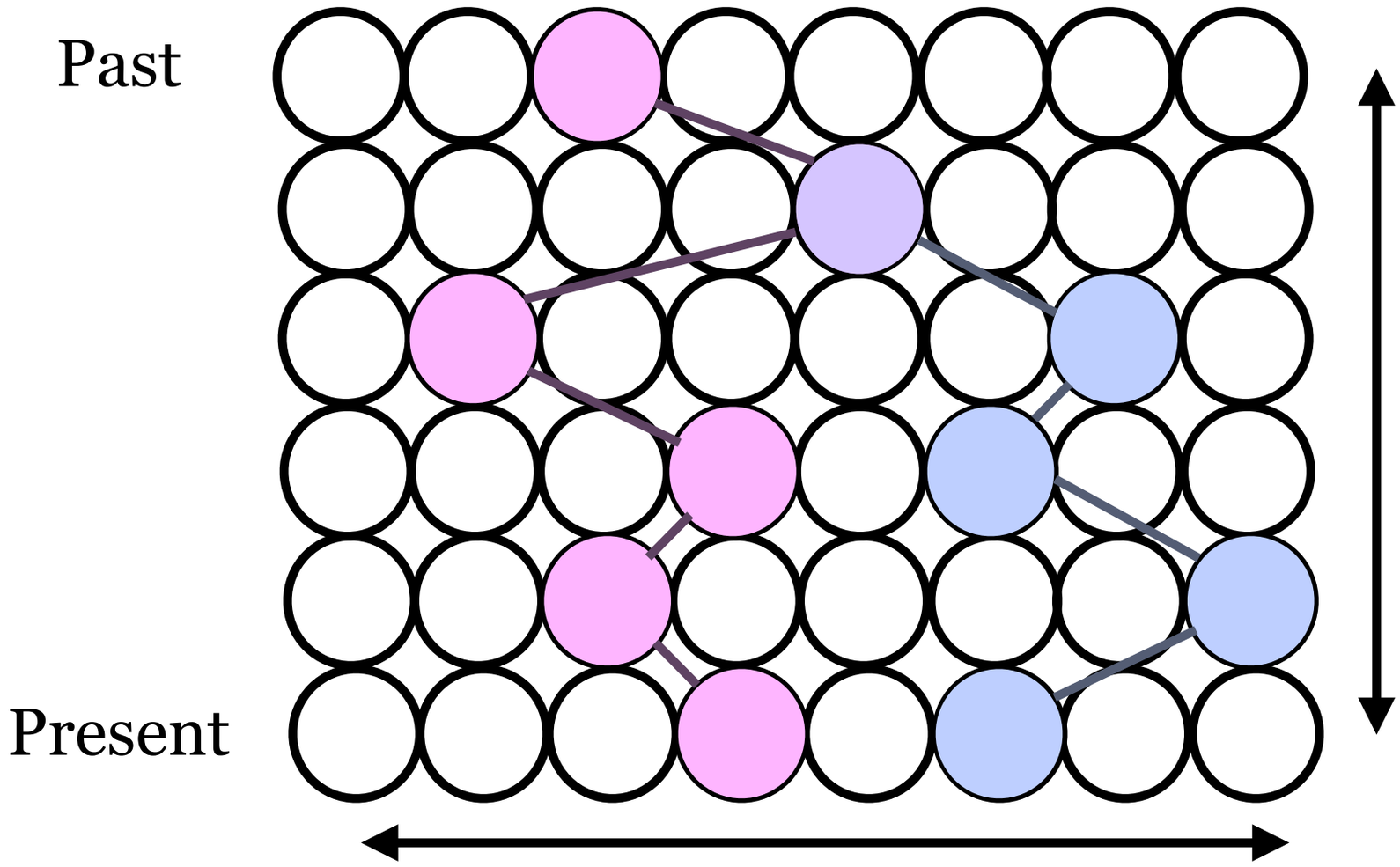


- Haploid population of constant size (N)
- No selection: uniform probability 1/N

Probability of no coalescence in g generations: $\left(1-\frac{1}{N}\right)^{g}$

$$t = g/N \Rightarrow \left(1-\frac{t}{Nt}\right)^{Nt} \xrightarrow[N\to\infty]{} e^{-t}$$

Coalescent model within 1 population



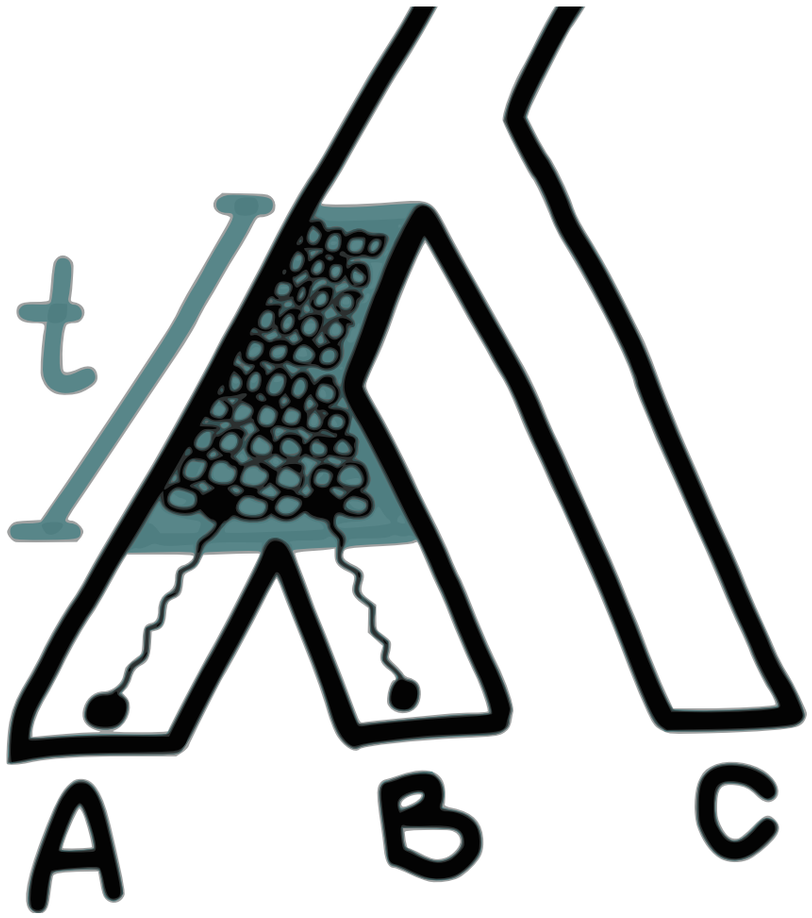
- Haploid population of constant size (N)
- No selection: uniform probability 1/N

Probability of no coalescence in g generations: $\left(1-\frac{1}{N}\right)^{g}$

$$t = g/N \Rightarrow \left(1-\frac{t}{Nt}\right)^{Nt} \xrightarrow[N\to\infty]{} e^{-t}$$

2N

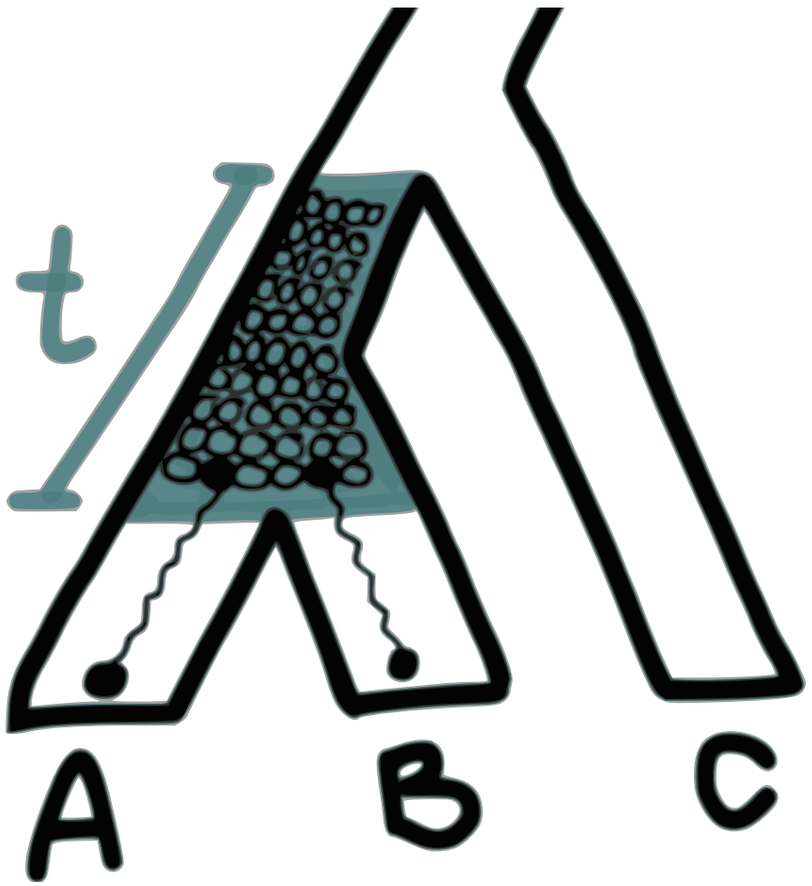
Multispecies coalescent on a tree



$$P(T > t) = e^{-t}$$

$$T = \frac{g}{N} \text{ coalescent units } \sim Exp(1)$$

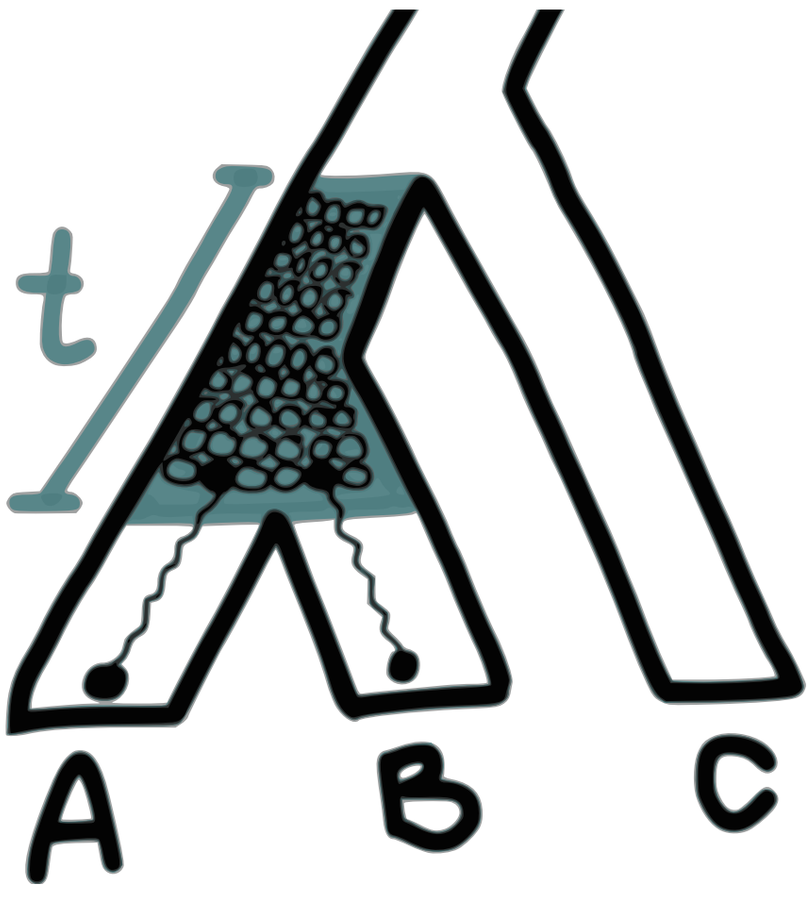
Multispecies coalescent on a tree



$$\mathbb{P}\left(\begin{array}{ccc} & \wedge & \\ A & B & C \end{array}\right) =$$

$$P(T > t) = e^{-t}$$

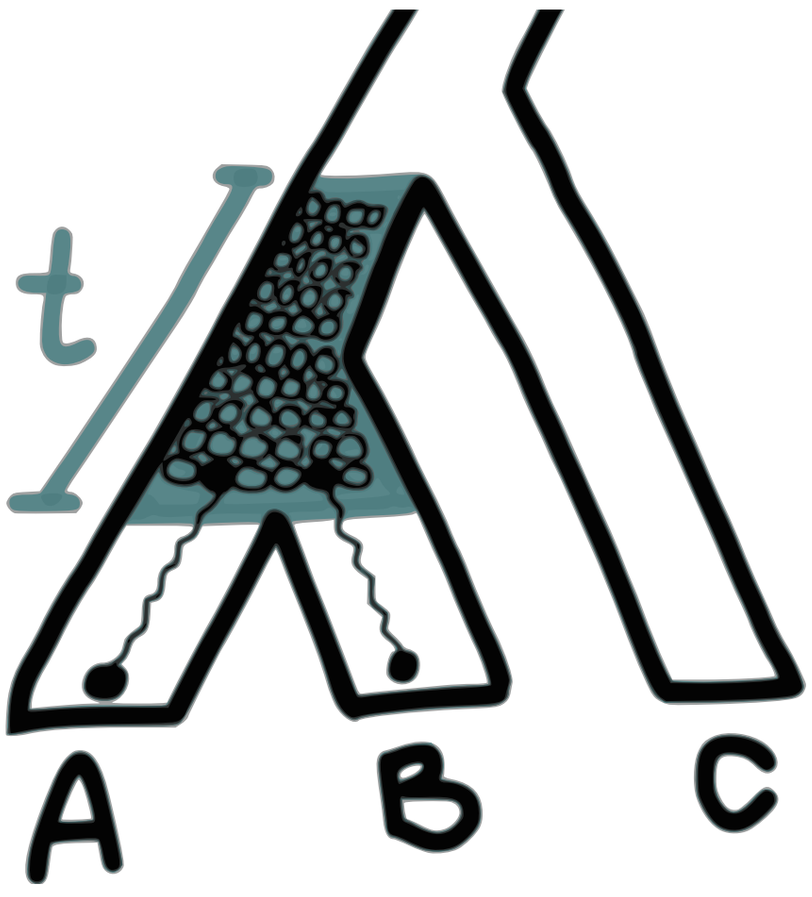
Multispecies coalescent on a tree



$$P(T > t) = e^{-t}$$

$$\mathbb{P}\left(\begin{array}{ccc} & \wedge & \\ A & B & C \end{array}\right) = 1 - e^{-t}$$

Multispecies coalescent on a tree



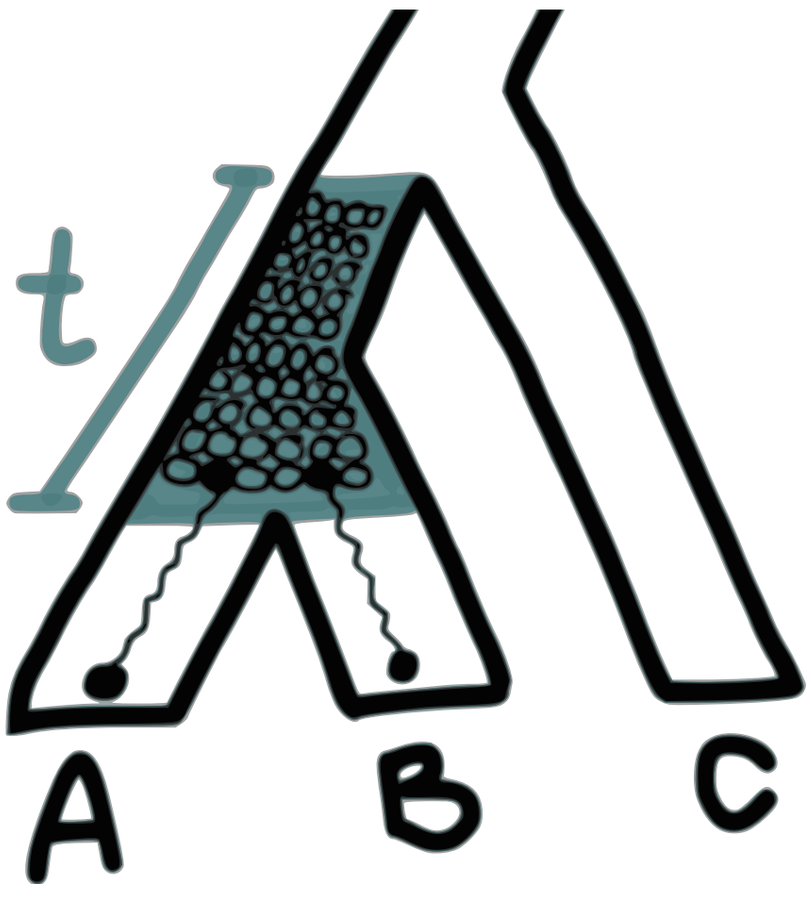
$$P(T > t) = e^{-t}$$

$$\mathbb{P}\left(\begin{array}{ccc} & \wedge & \\ A & B & C \end{array}\right) =$$

$$1 - e^{-t}$$

$$+$$

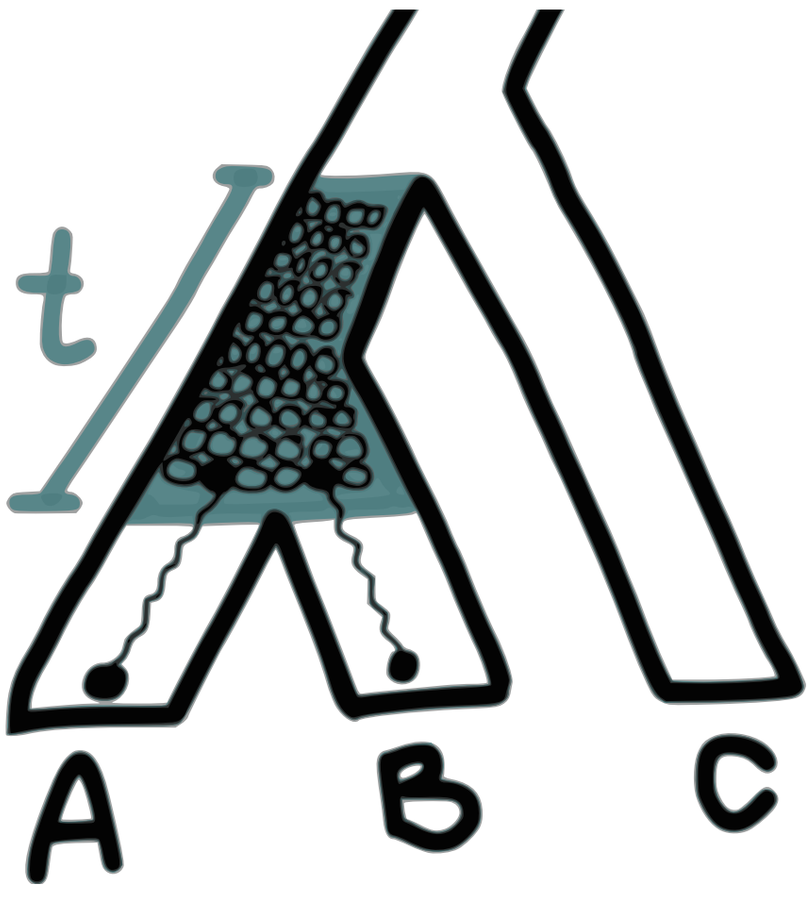
Multispecies coalescent on a tree



$$\mathbb{P}\left(\begin{array}{ccc} & \wedge & \\ A & B & C \end{array}\right) = 1 - e^{-t} + e^{-t} \times 1/3$$

$$P(T > t) = e^{-t}$$

Multispecies coalescent on a tree



$$P(T > t) = e^{-t}$$

$$\mathbb{P}\left(\begin{array}{ccc} & \wedge & \\ A & B & C \end{array}\right) =$$

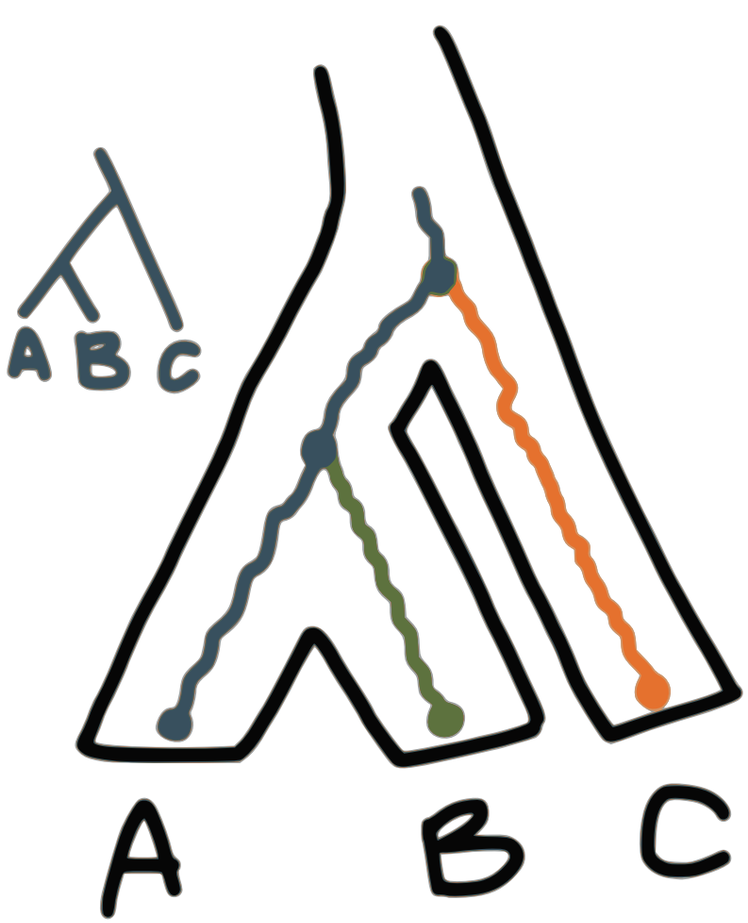
$$1 - e^{-t}$$

$$+$$

$$e^{-t} \times 1/3$$

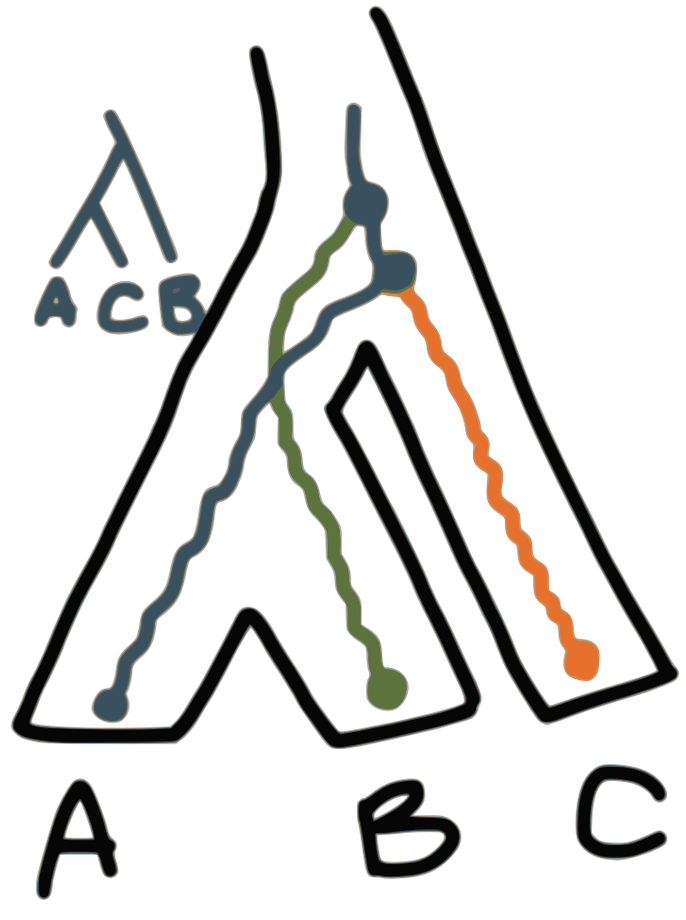
$$= 1 - \frac{2}{3} e^{-t}$$

Multispecies coalescent on a tree



$$1 - \frac{2}{3}e^{-t}$$

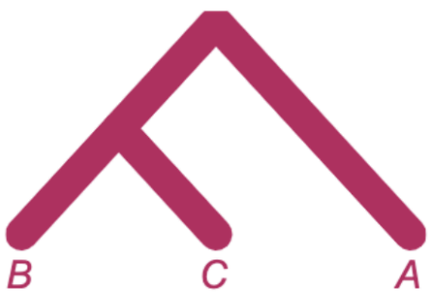
$$\frac{1}{3}e^{-t}$$



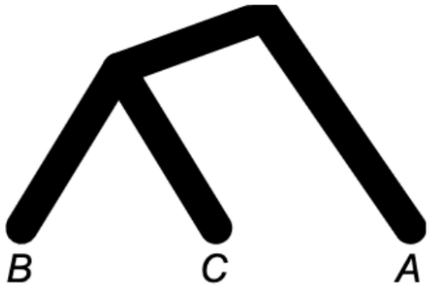
$$\frac{1}{3}e^{-t}$$

Phylogenetic coalescent model

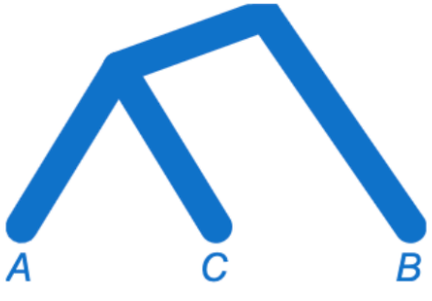
Probabilities of each gene tree history are shown below them
t = length of interval between speciation events



$1 - e^{-t}$



$\frac{1}{3}e^{-t}$

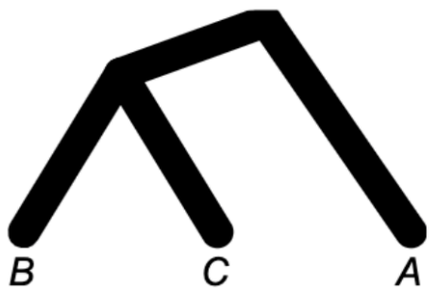


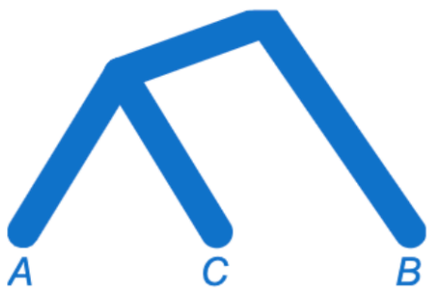
$\frac{1}{3}e^{-t}$

$\frac{1}{3}e^{-t}$

Phylogenetic coalescent model

t = length of interval between coalescent events $= 1.0 = 0.5 = 2.0$

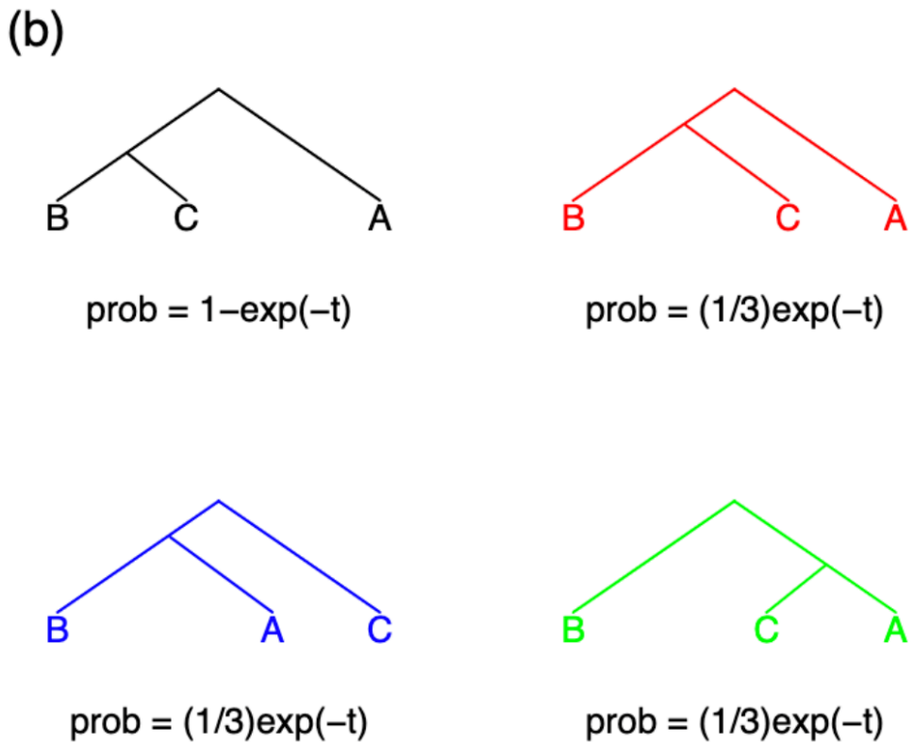


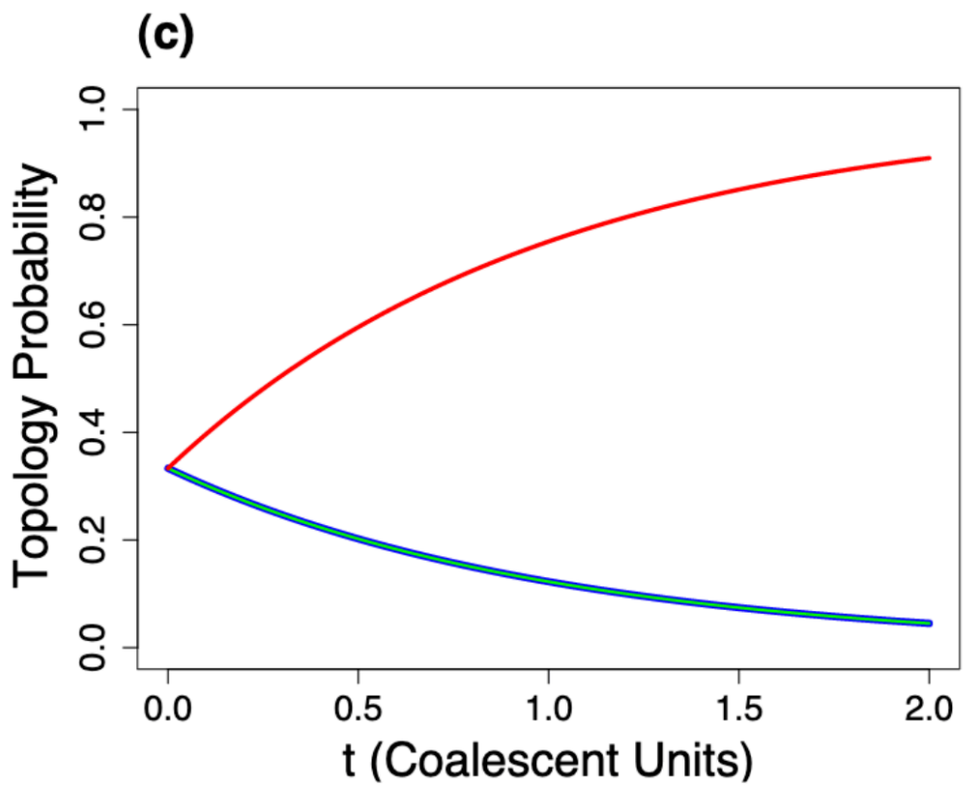


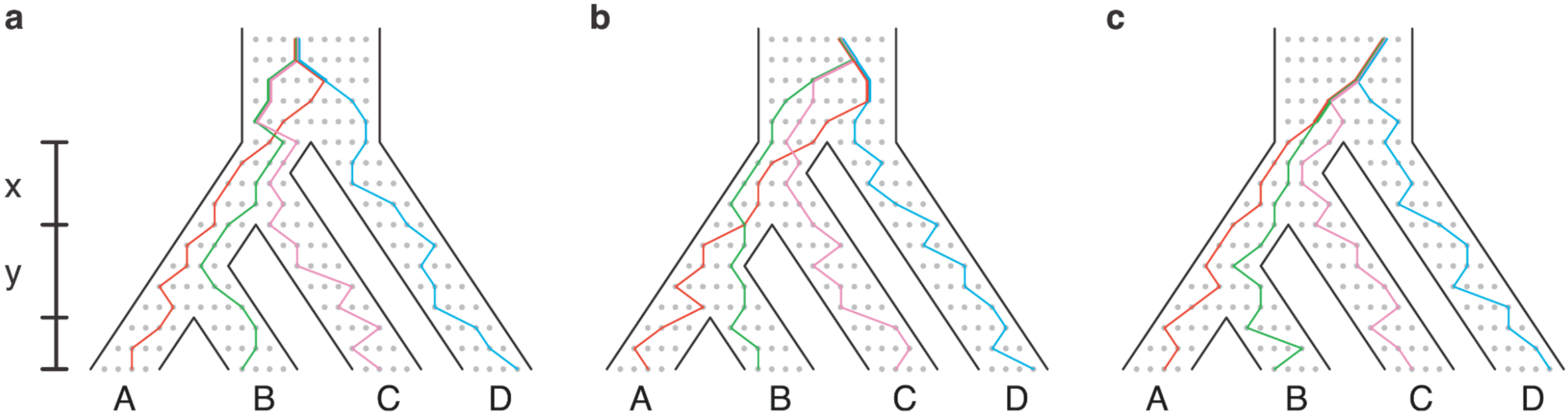
$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
0.40	0.20	0.20	0.20
0.85	0.05	0.05	0.05

Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

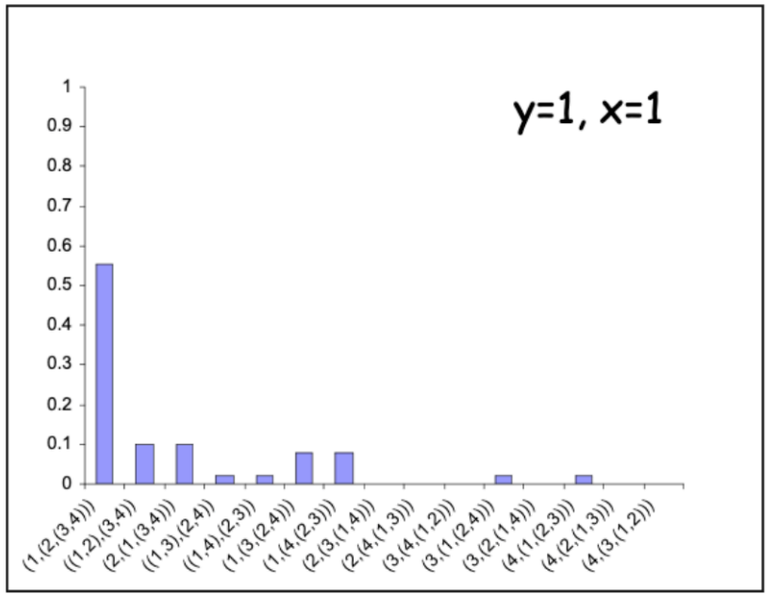
- What are these probabilities like as a function of t, the length of time between speciation events?

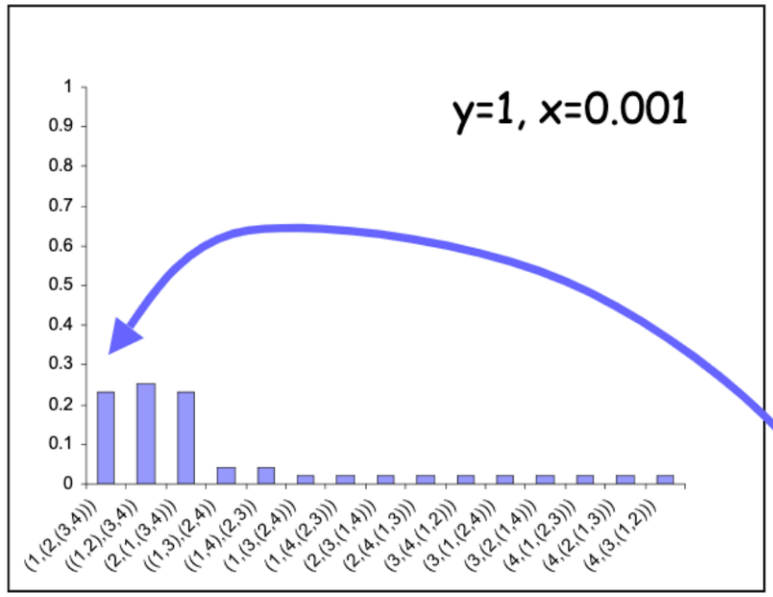


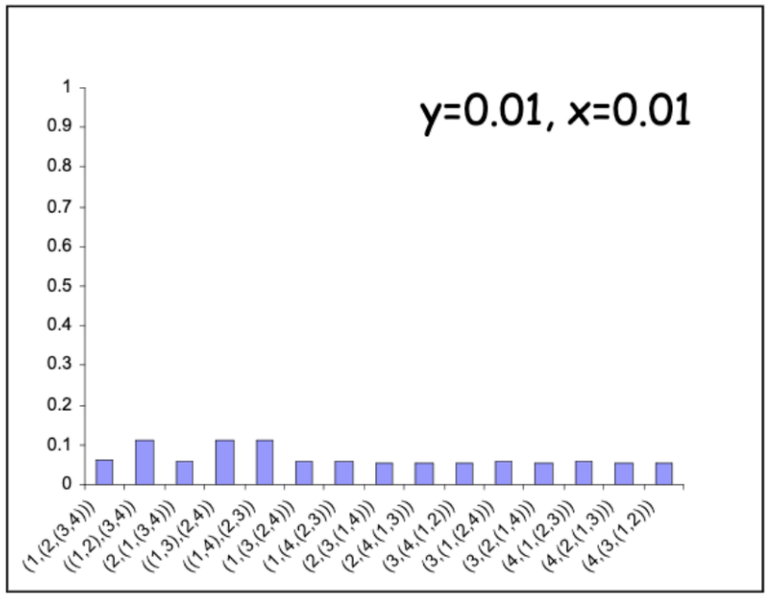


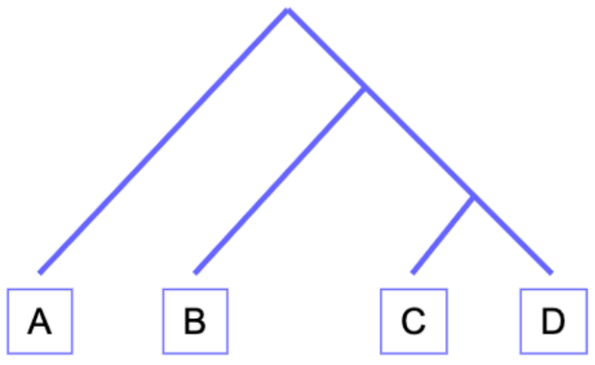


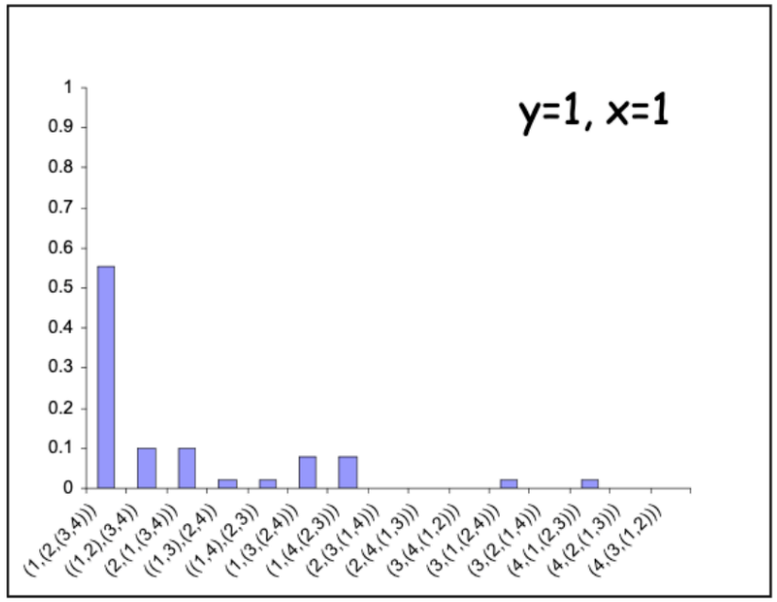
If the internal branches of the species tree—x and y—are short so that coalescences occur deep in the tree, the two sequences of coalescences that produce a given symmetric gene tree topology together have higher probability than the single sequence that produces the topology that matches the species tree. (a) and (b) Two coalescence sequences leading to gene tree topology ((AD)(BC)). In (a), the lineages from B and C coalesce more recently than those from A and D, and in (b), the reverse is true. (c) The single sequence of coalescences leading to gene tree topology (((AB)C)D). (Degnan and Rosenberg, 2006)

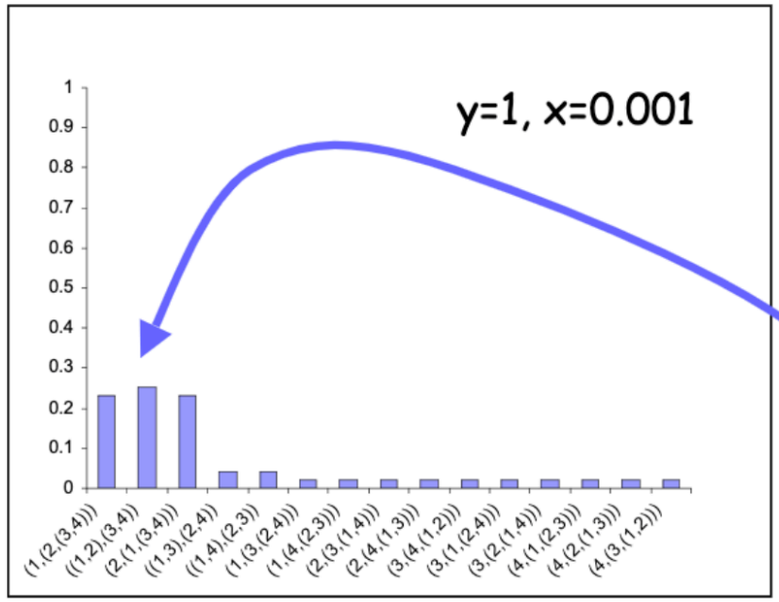

y=1, x=1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(1,(2,(3,4)))
((1,2),(3,4))
(2,(1,(3,4)))
((1,3),(2,4))
((1,4),(2,3))
(1,(3,(2,4)))
(1,(4,(2,3)))
(2,(3,(1,4)))
(2,(4,(1,3)))
(3,(4,(1,2)))
(3,(1,(2,4)))
(3,(2,(1,4)))
(4,(1,(2,3)))
(4,(2,(1,3)))
(4,(3,(1,2)))

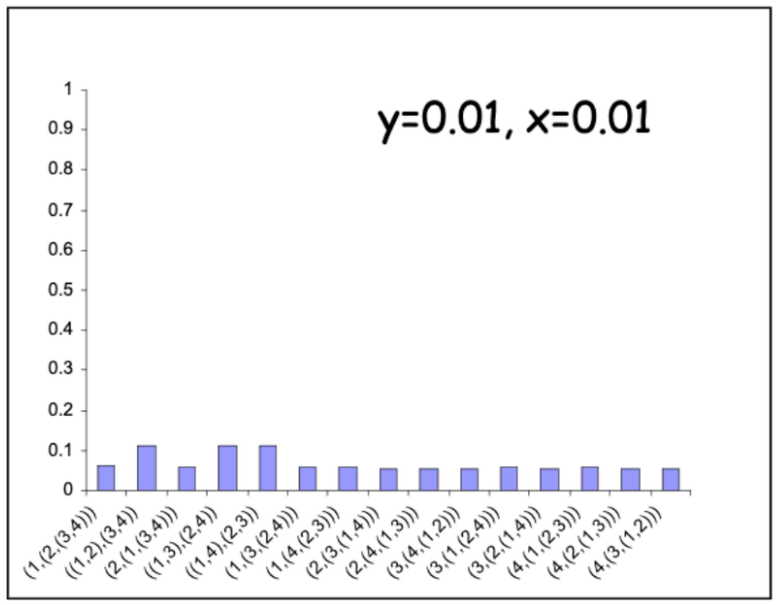

y=1, x=0.001
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(1,(2,(3,4)))
((1,2),(3,4))
(2,(1,(3,4)))
((1,3),(2,4))
((1,4),(2,3))
(1,(3,(2,4)))
(1,(4,(2,3)))
(2,(3,(1,4)))
(2,(4,(1,3)))
(3,(4,(1,2)))
(3,(1,(2,4)))
(3,(2,(1,4)))
(4,(1,(2,3)))
(4,(2,(1,3)))
(4,(3,(1,2)))

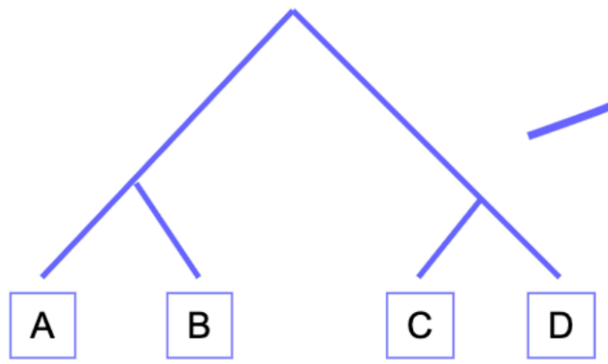

y=0.01, x=0.01
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(1,(2,(3,4)))
((1,2),(3,4))
(2,(1,(3,4)))
((1,3),(2,4))
((1,4),(2,3))
(1,(3,(2,4)))
(1,(4,(2,3)))
(2,(3,(1,4)))
(2,(4,(1,3)))
(3,(4,(1,2)))
(3,(1,(2,4)))
(3,(2,(1,4)))
(4,(1,(2,3)))
(4,(2,(1,3)))
(4,(3,(1,2)))

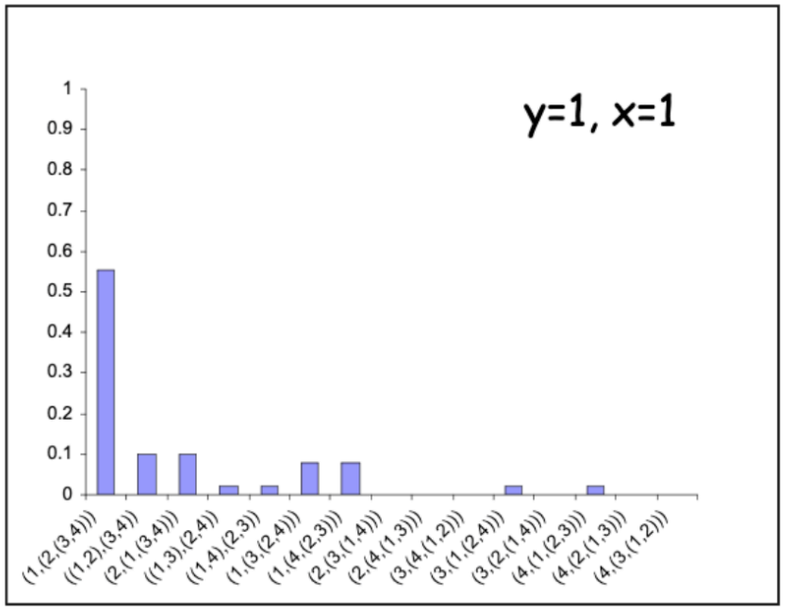

A
B
C
D

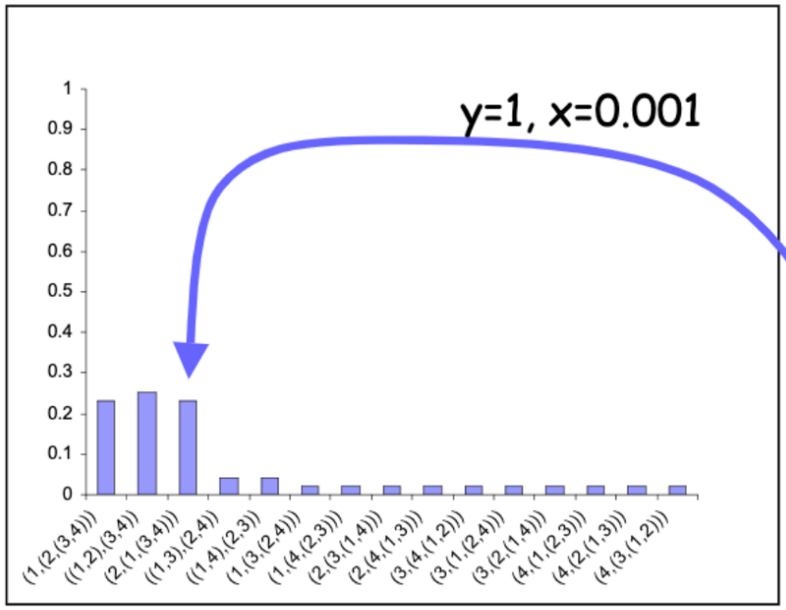

y=1, x=1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(1,(2,(3,4)))
((1,2),(3,4))
(2,(1,(3,4)))
((1,3),(2,4))
((1,4),(2,3))
(1,(3,(2,4)))
(1,(4,(2,3)))
(2,(3,(1,4)))
(2,(4,(1,3)))
(3,(4,(1,2)))
(3,(1,(2,4)))
(3,(2,(1,4)))
(4,(1,(2,3)))
(4,(2,(1,3)))
(4,(3,(1,2)))

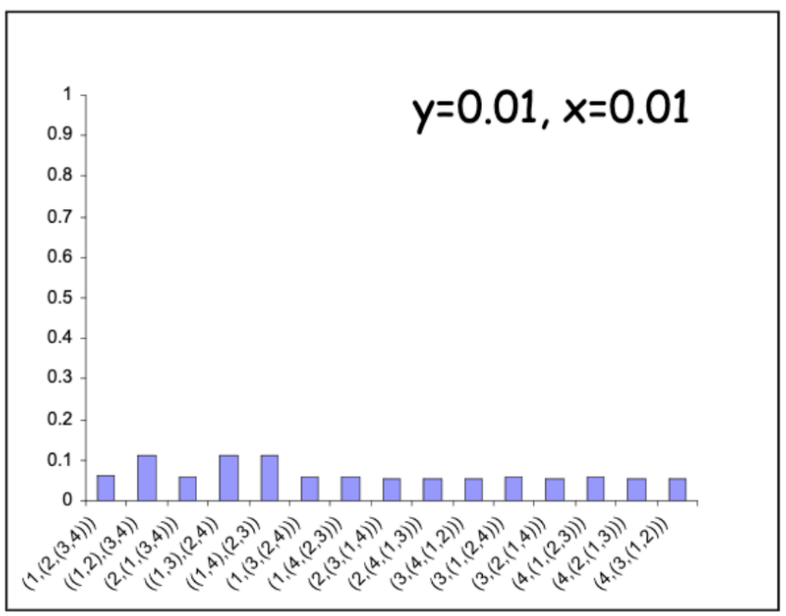

y=1, x=0.001
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(1,(2,(3,4)))
((1,2),(3,4))
(2,(1,(3,4)))
((1,3),(2,4))
((1,4),(2,3))
(1,(3,(2,4)))
(1,(4,(2,3)))
(2,(3,(1,4)))
(2,(4,(1,3)))
(3,(4,(1,2)))
(3,(1,(2,4)))
(3,(2,(1,4)))
(4,(1,(2,3)))
(4,(2,(1,3)))
(4,(3,(1,2)))

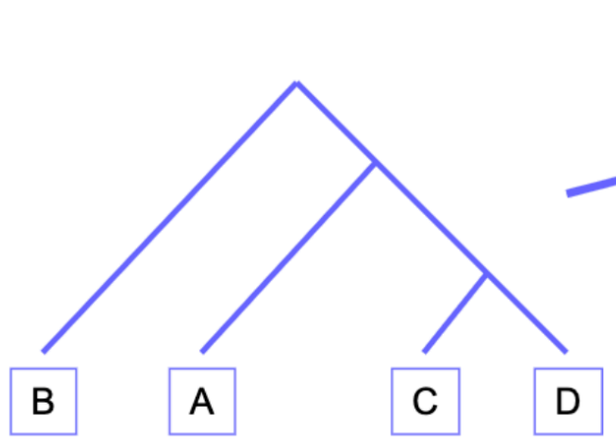

y=0.01, x=0.01
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(1,(2,(3,4)))
((1,2),(3,4))
(2,(1,(3,4)))
((1,3),(2,4))
((1,4),(2,3))
(1,(3,(2,4)))
(1,(4,(2,3)))
(2,(3,(1,4)))
(2,(4,(1,3)))
(3,(4,(1,2)))
(3,(1,(2,4)))
(3,(2,(1,4)))
(4,(1,(2,3)))
(4,(2,(1,3)))
(4,(3,(1,2)))


A
B
C
D


y=1, x=1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(1,(2,(3,4)))
((1,2),(3,4))
(2,(1,(3,4)))
((1,3),(2,4))
((1,4),(2,3))
(1,(3,(2,4)))
(1,(4,(2,3)))
(2,(3,(1,4)))
(2,(4,(1,3)))
(3,(4,(1,2)))
(3,(1,(2,4)))
(3,(2,(1,4)))
(4,(1,(2,3)))
(4,(2,(1,3)))
(4,(3,(1,2)))


y=1, x=0.001
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(1,(2,(3,4)))
((1,2),(3,4))
(2,(1,(3,4)))
((1,3),(2,4))
((1,4),(2,3))
(1,(3,(2,4)))
(1,(4,(2,3)))
(2,(3,(1,4)))
(2,(4,(1,3)))
(3,(4,(1,2)))
(3,(1,(2,4)))
(3,(2,(1,4)))
(4,(1,(2,3)))
(4,(2,(1,3)))
(4,(3,(1,2)))


y=0.01, x=0.01
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(1,(2,(3,4)))
((1,2),(3,4))
(2,(1,(3,4)))
((1,3),(2,4))
((1,4),(2,3))
(1,(3,(2,4)))
(1,(4,(2,3)))
(2,(3,(1,4)))
(2,(4,(1,3)))
(3,(4,(1,2)))
(3,(1,(2,4)))
(3,(2,(1,4)))
(4,(1,(2,3)))
(4,(2,(1,3)))
(4,(3,(1,2)))


B
A
C
D

Anomaly zone

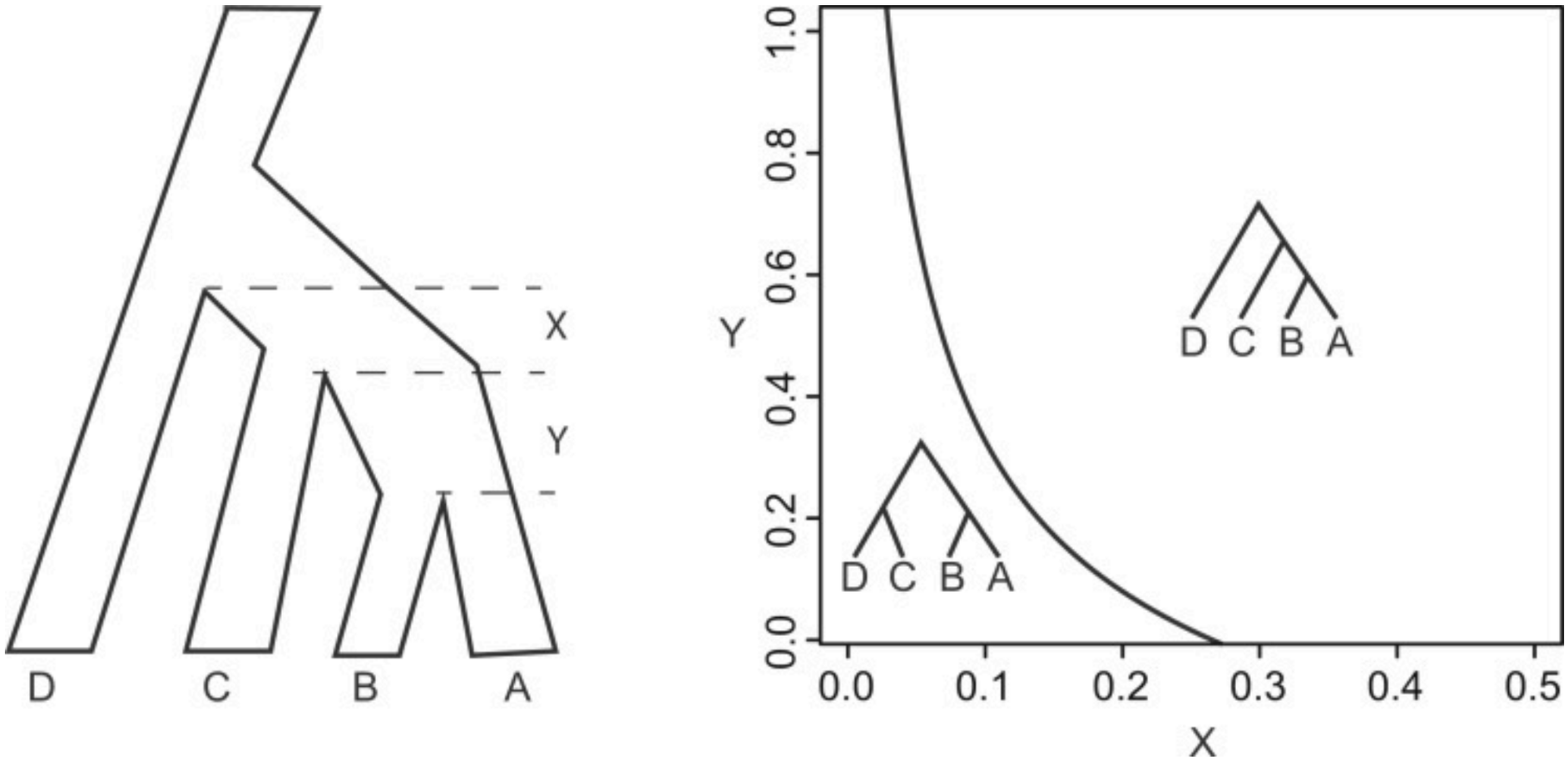
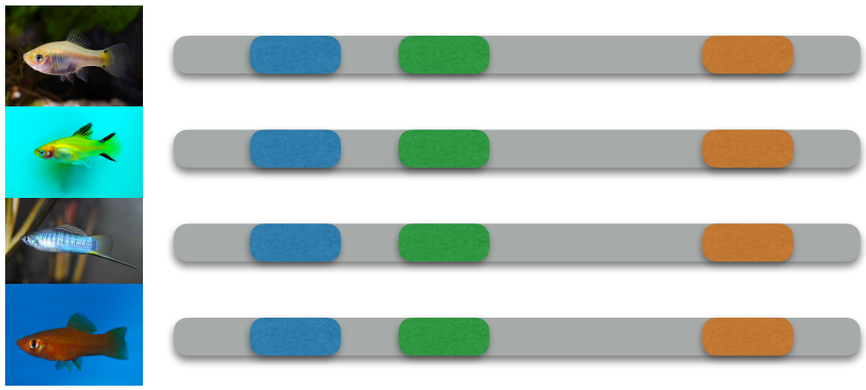


Figure 1 in Linkem et al (2016)



Distances
Parsimony
Likelihood
(Bayesian)

$$P(T,\theta|G) \propto \pi(T)\pi(\theta)\prod_{i=1}^{L} P(G_i|T,\theta)$$

Bayesian

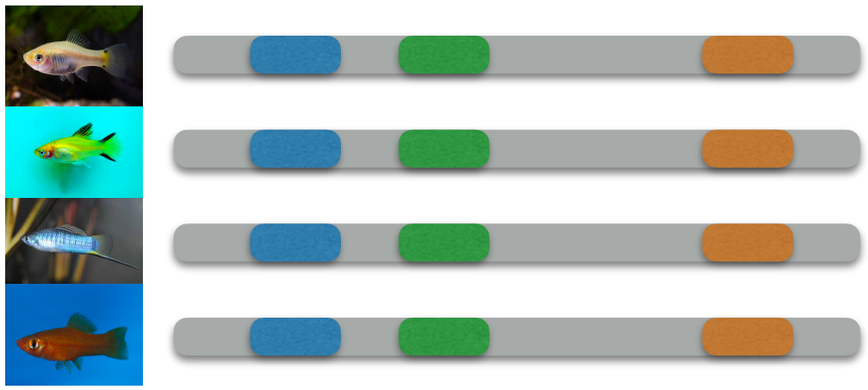
Prior Tree — **Multispecies Coalescent**

$$L(T,\theta) = \prod_{i=1}^{L} P(G_i|T,\theta)$$

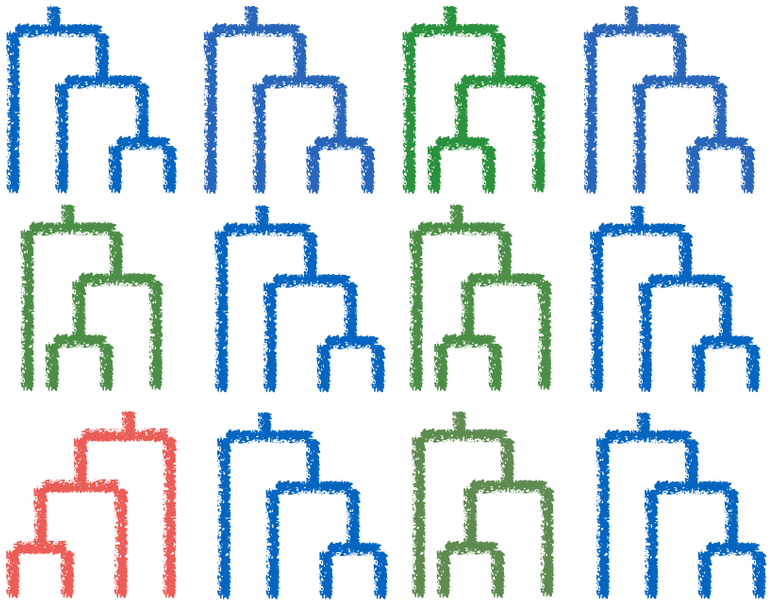
Max. Lik.

Summary methods: ASTRAL, BUCKy, MP-EST

(Zhang et al, 2018) (Larget et al, 2010) (Liu et al, 2010)



Distances
Parsimony
Likelihood
(Bayesian)



$$P(T,\theta|G) \propto \pi(T)\pi(\theta)\prod_{i=1}^{L} P(G_i|T,\theta)$$

Prior
Tree

Multispecies
Coalescent

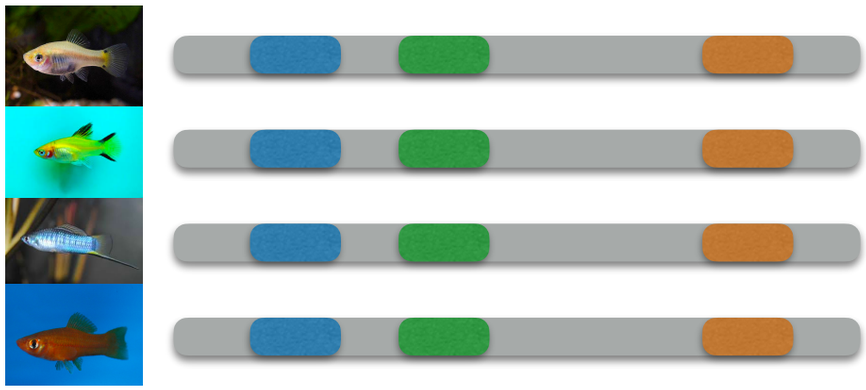
Bayesian

$$L(T,\theta) = \prod_{i=1}^{L} P(G_i|T,\theta)$$

Max. Lik.

Summary methods: ASTRAL, BUCKy, MP-EST

(Zhang et al, 2018) (Larget et al, 2010) (Liu et al, 2010)



Distances
Parsimony
Likelihood
(Bayesian)

$$P(T,\theta|G) \propto \pi(T)\pi(\theta)\prod_{i=1}^{L} P(G_i|T,\theta)$$

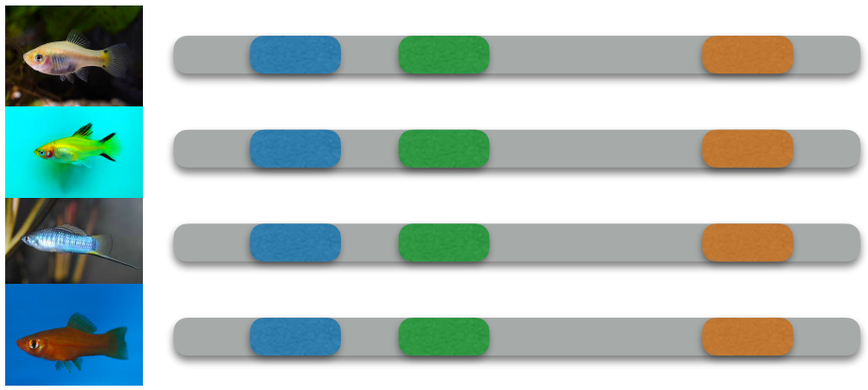
Bayesian

Prior Tree — **Multispecies Coalescent**

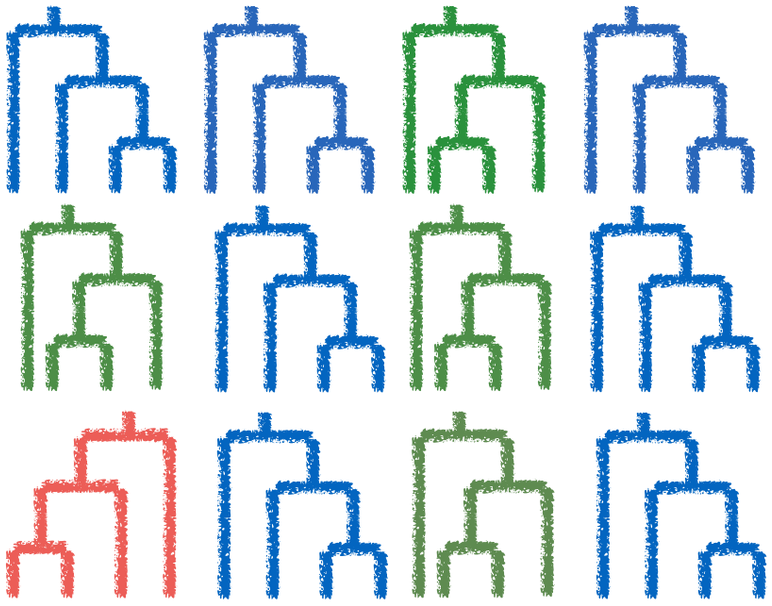
$$L(T,\theta) = \prod_{i=1}^{L} P(G_i|T,\theta)$$

Max. Lik.

Summary methods: Do not search tree space ASTRAL, BUCKy, MP-EST
(Zhang et al, 2018) (Larget et al, 2010) (Liu et al, 2010)



Distances
Parsimony
Likelihood
(Bayesian)



$$P(T,\theta|G) \propto \pi(T)\pi(\theta)\prod_{i=1}^{L} P(G_i|T,\theta)$$

Prior Tree

Multispecies Coalescent

Bayesian

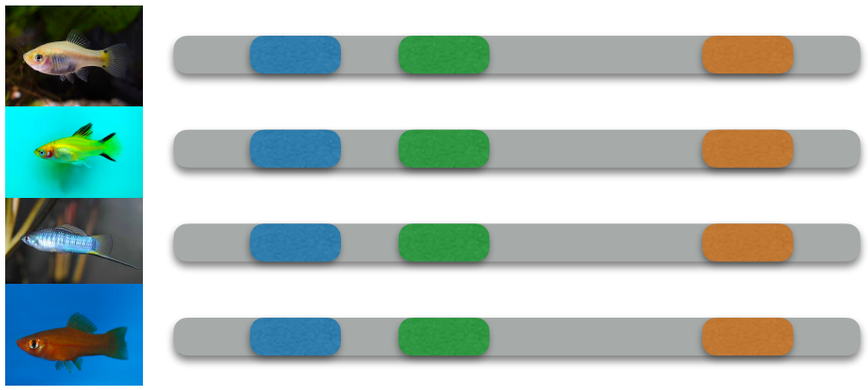
$$L(T,\theta) = \prod_{i=1}^{L} P(G_i|T,\theta)$$

Max. Lik.

Summary methods: Do not search tree space

ASTRAL, BUCKy, MP-EST

(Zhang et al, 2018) (Larget et al, 2010) (Liu et al, 2010)



Co-estimation (lecture 15)

$$P(T, G, \theta | D) \propto \pi(T)\pi(\theta) \prod_{i=1}^{L} P(D_i|G_i)P(G_i|T)$$

Tree — gene trees — L loci — Prior Tree — Substitution Model — Multispecies Coalescent

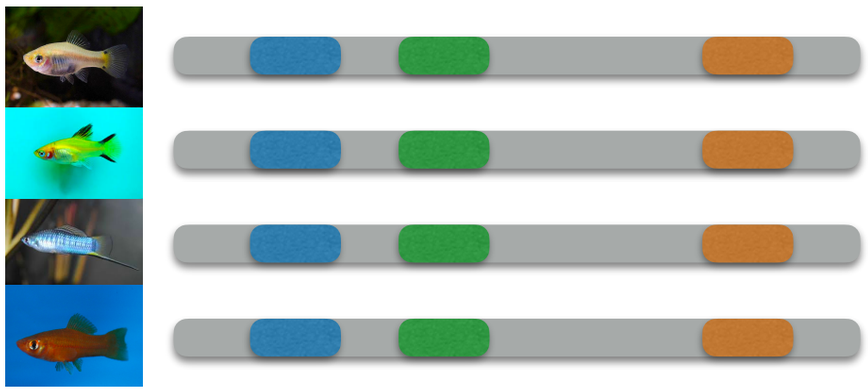
$$P(T, \theta | G) \propto \pi(T)\pi(\theta) \prod_{i=1}^{L} P(G_i|T, \theta)$$ Bayesian

$$L(T, \theta) = \prod_{i=1}^{L} P(G_i|T, \theta)$$ Max. Lik.

Summary methods: ASTRAL, BUCKy, MP-EST

(Zhang et al, 2018) (Larget et al, 2010) (Liu et al, 2010)

Co-estimation (lecture 15)



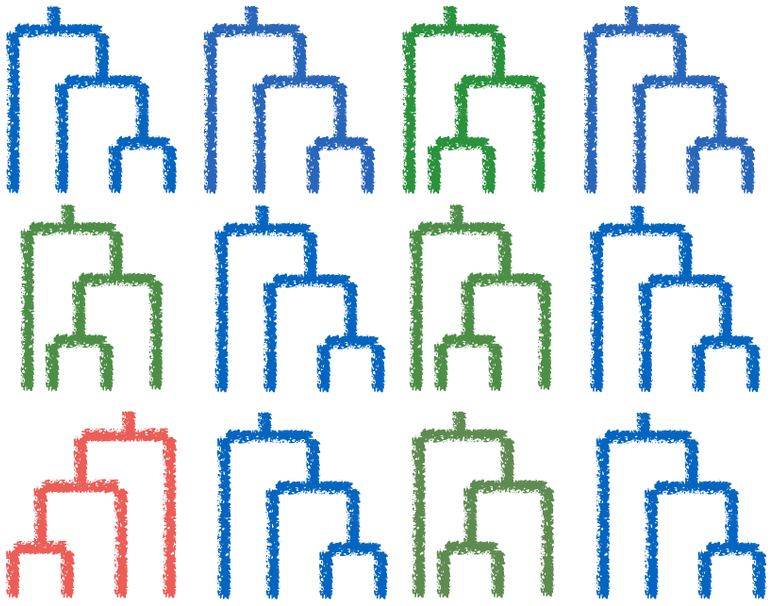
$$P(T, G, \theta | D) \propto \pi(T)\pi(\theta) \prod_{i=1}^{L} P(D_i|G_i) P(G_i|T)$$

Tree — gene trees — L loci

Prior Tree

Substitution Model

Multispecies Coalescent



$$P(T, \theta | G) \propto \pi(T)\pi(\theta) \prod_{i=1}^{L} P(G_i|T, \theta)$$

Bayesian

$$L(T, \theta) = \prod_{i=1}^{L} P(G_i|T, \theta)$$

Max. Lik.

Summary methods: ASTRAL, BUCKy, MP-EST

(Zhang et al, 2018) (Larget et al, 2010) (Liu et al, 2010)

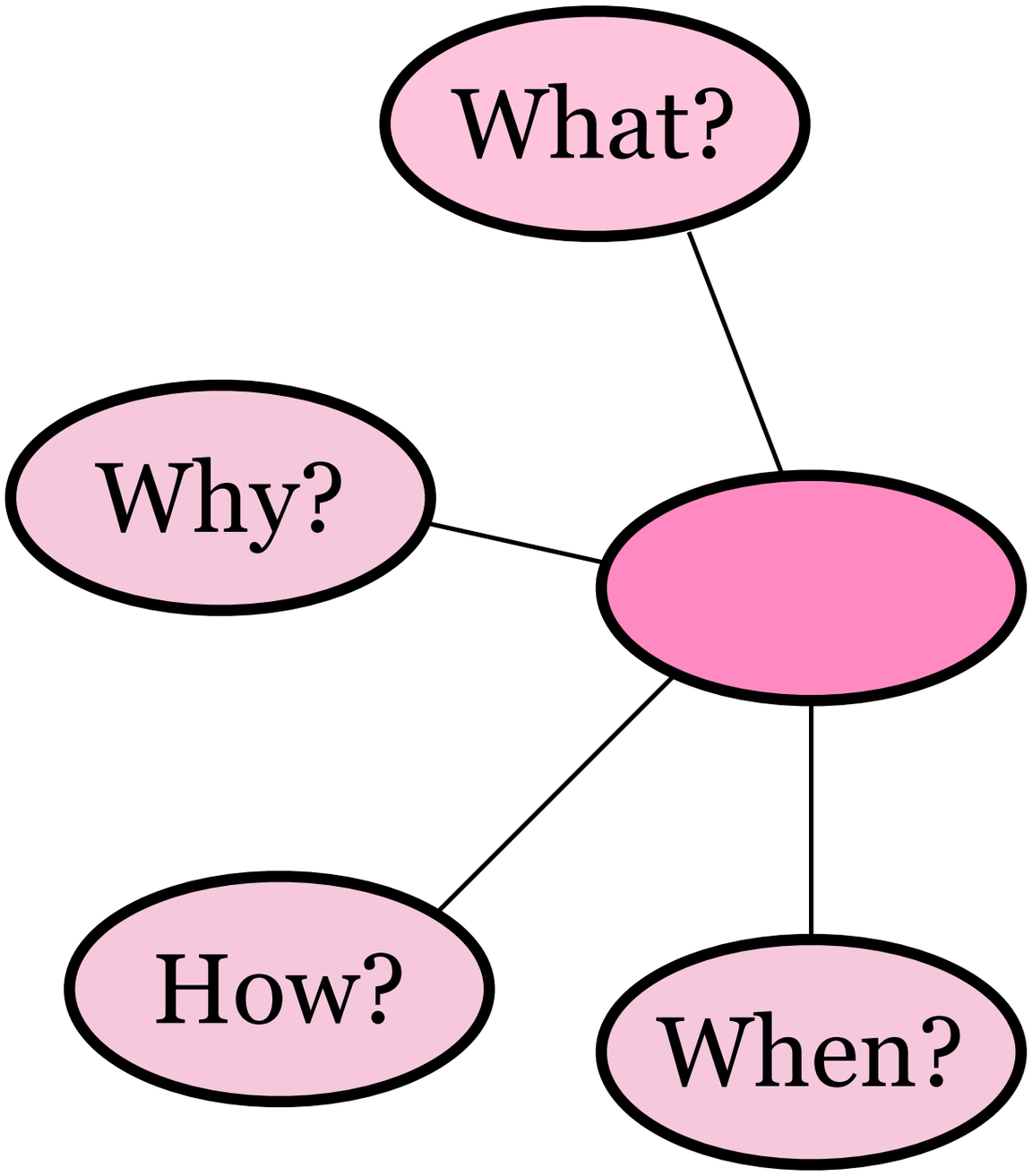
- Concatenation: assumes all genes follow the same tree-like history

- Coalescent-based tree methods: accounts for ILS (and sometimes gene tree estimation error), but does not allow for other sources of gene tree discordance (like GDL or gene flow)

- Extensions to coalescent-based tree methods:

    - Coalescent-based network methods: accounts for ILS, gene tree estimation error and gene flow

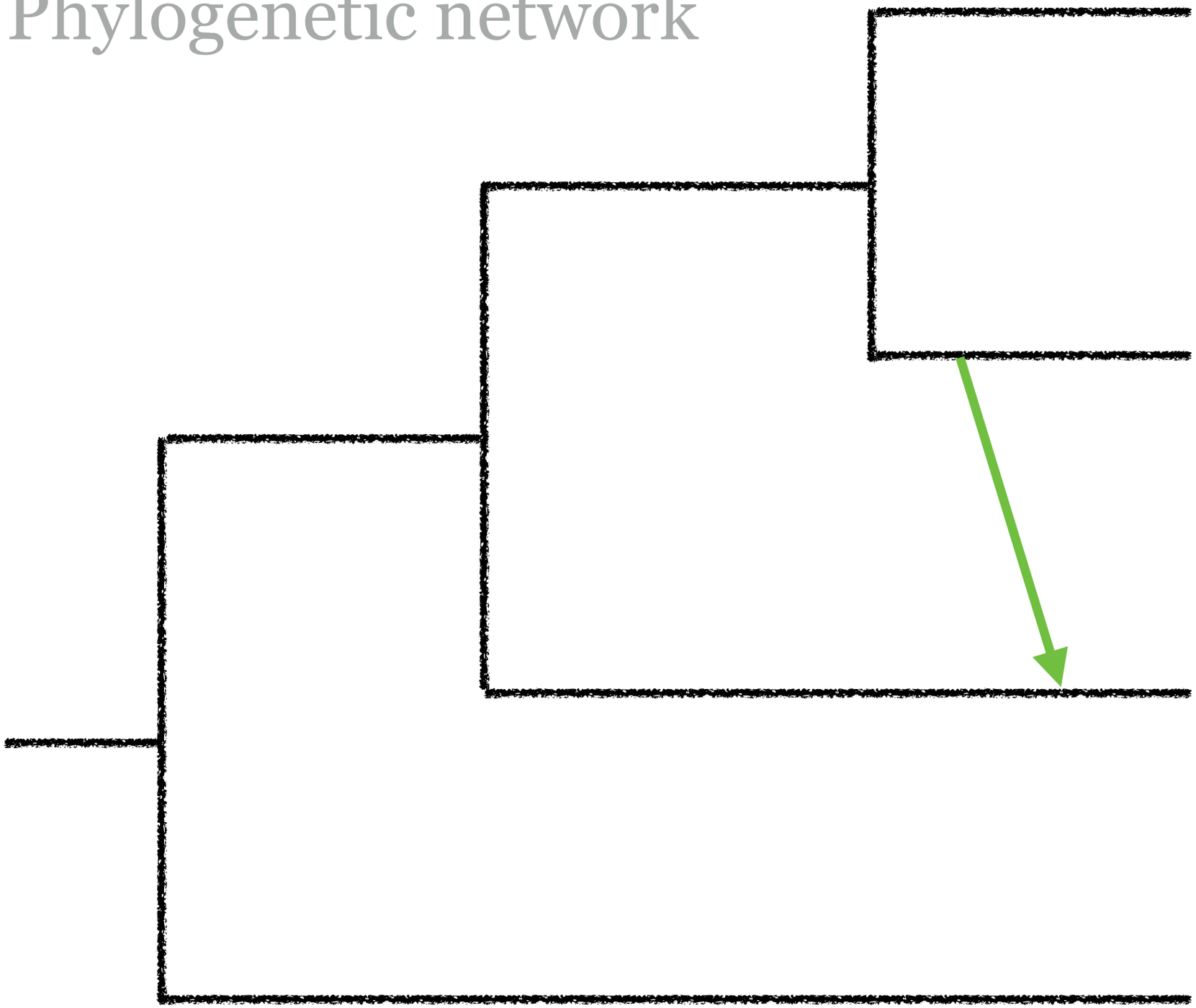
    - Coalescent+GDL methods (Li et al, 2020)



Phylogenetic Networks

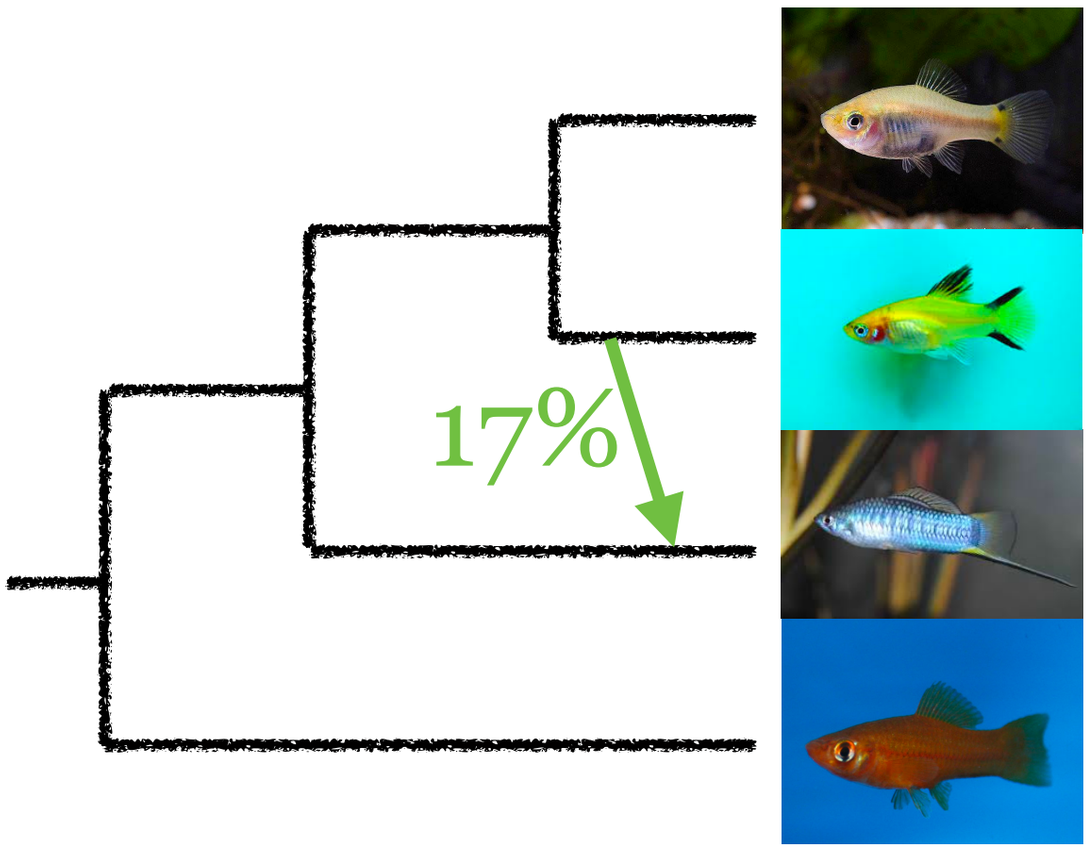
What?

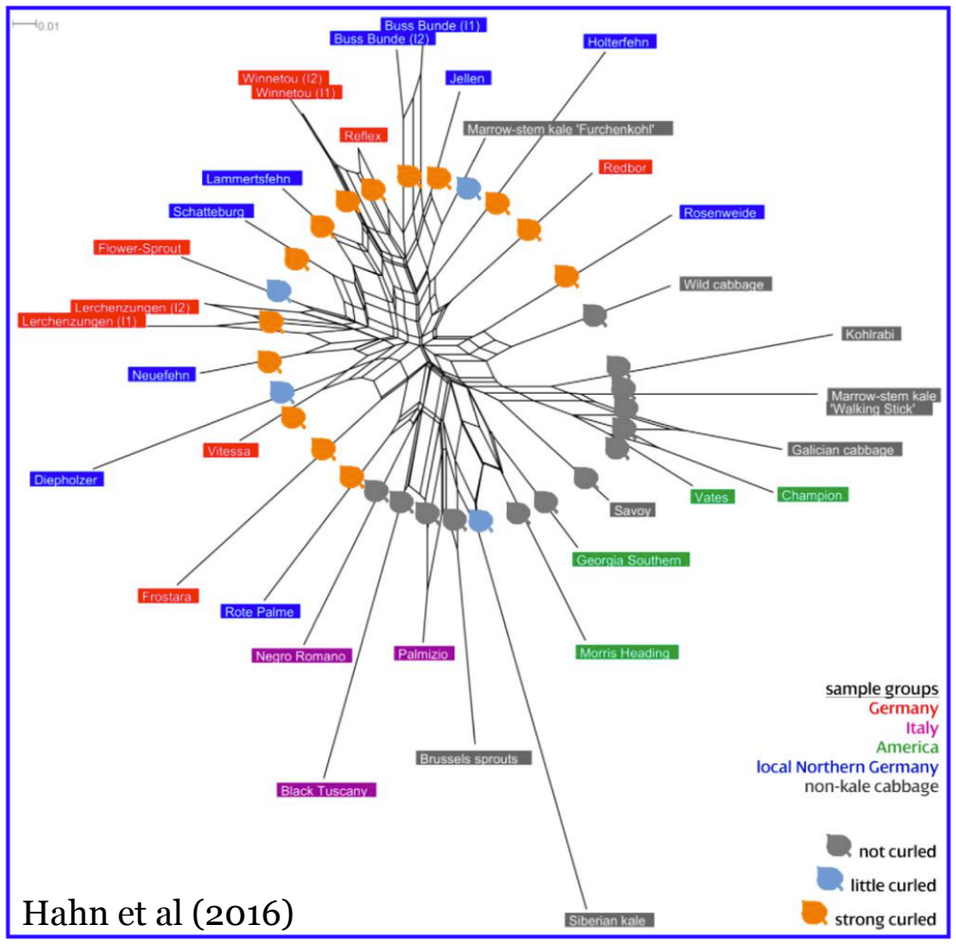
Phylogenetic network



What?

Phylogenetic network





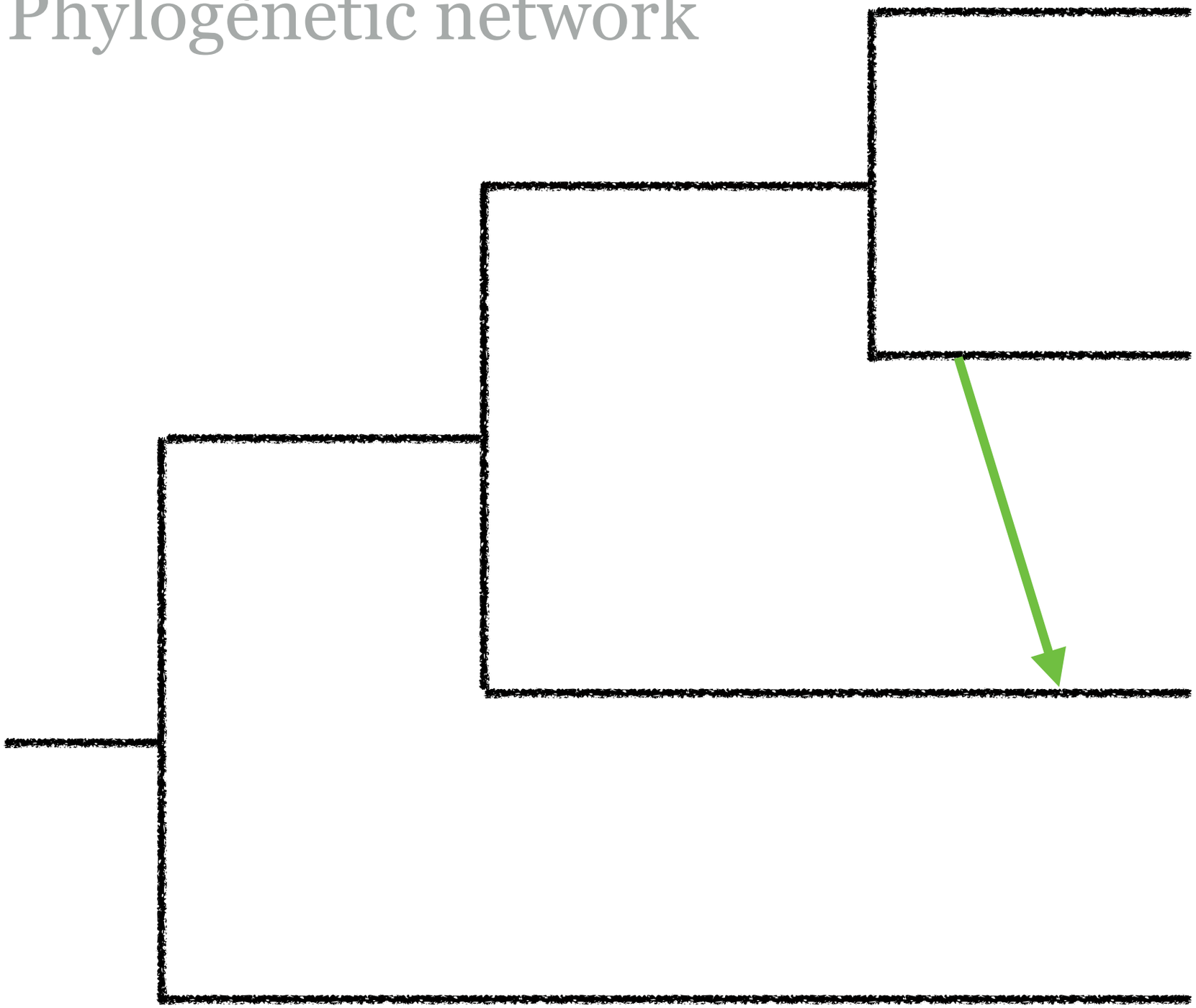
Hahn et al (2016)

Explicit

Implicit

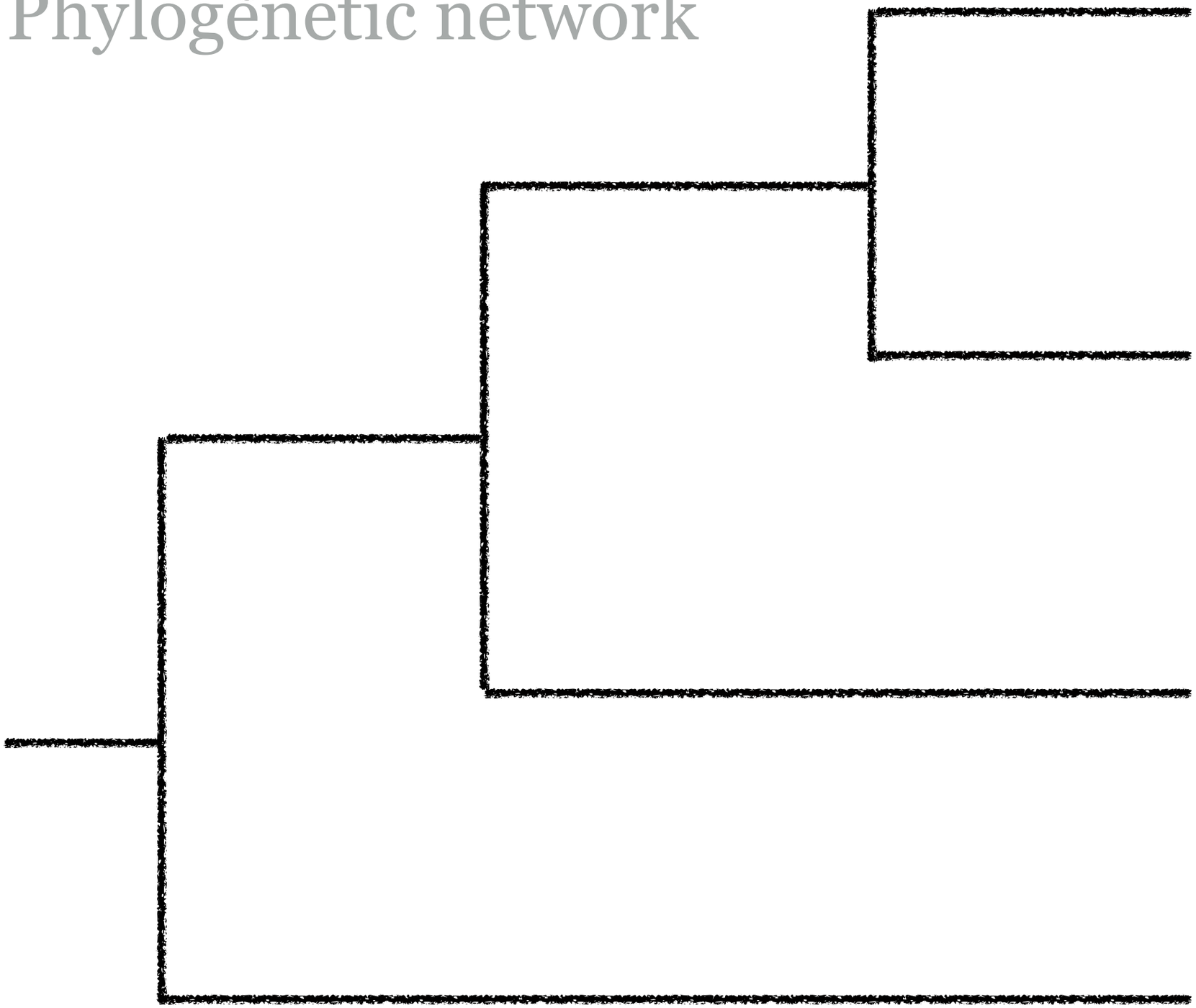
Why?

Phylogenetic network



Why?

Phylogenetic network

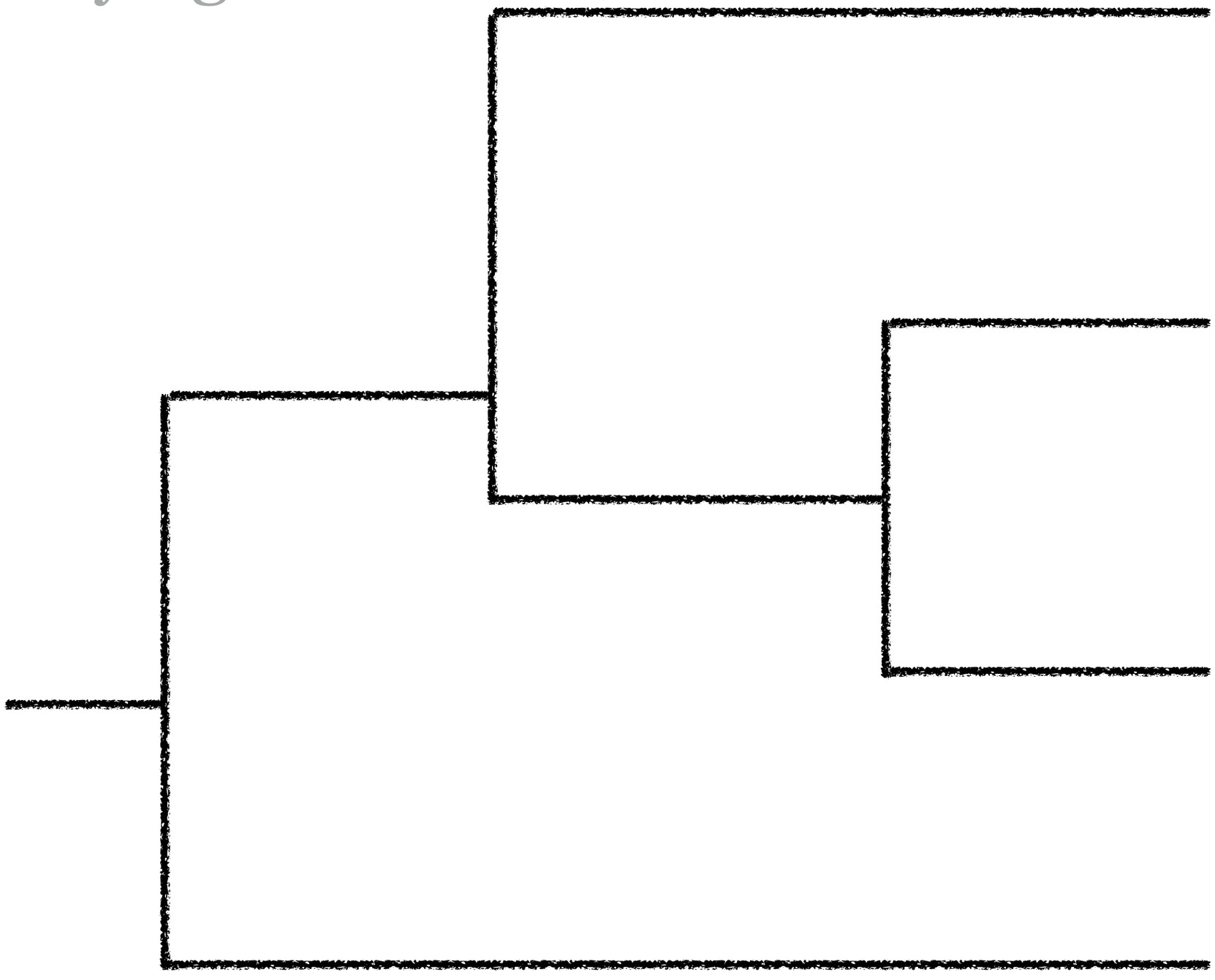


Main tree

Why?

Phylogenetic network

Ignore gene flow
=>Wrong tree!

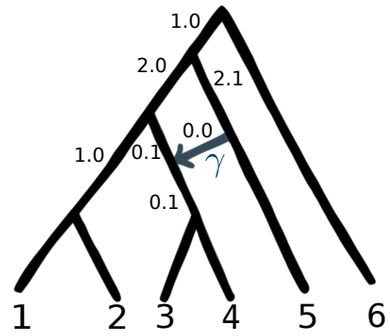


(S.-L., Yang, Ané, 2016, Syst Bio)

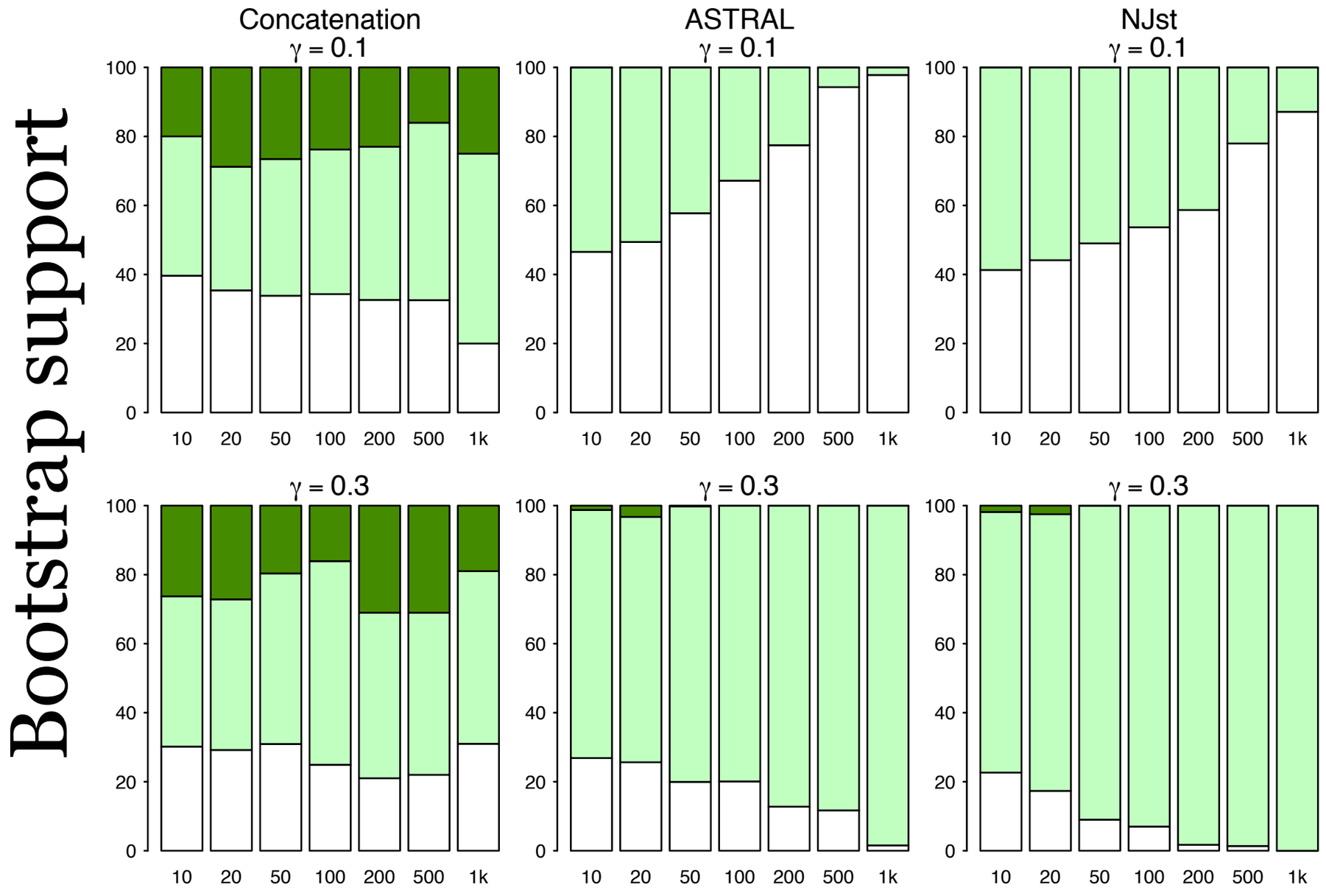
Why?

Phylogenetic network

Coalescent tree methods not robust to gene flow



White: true tree



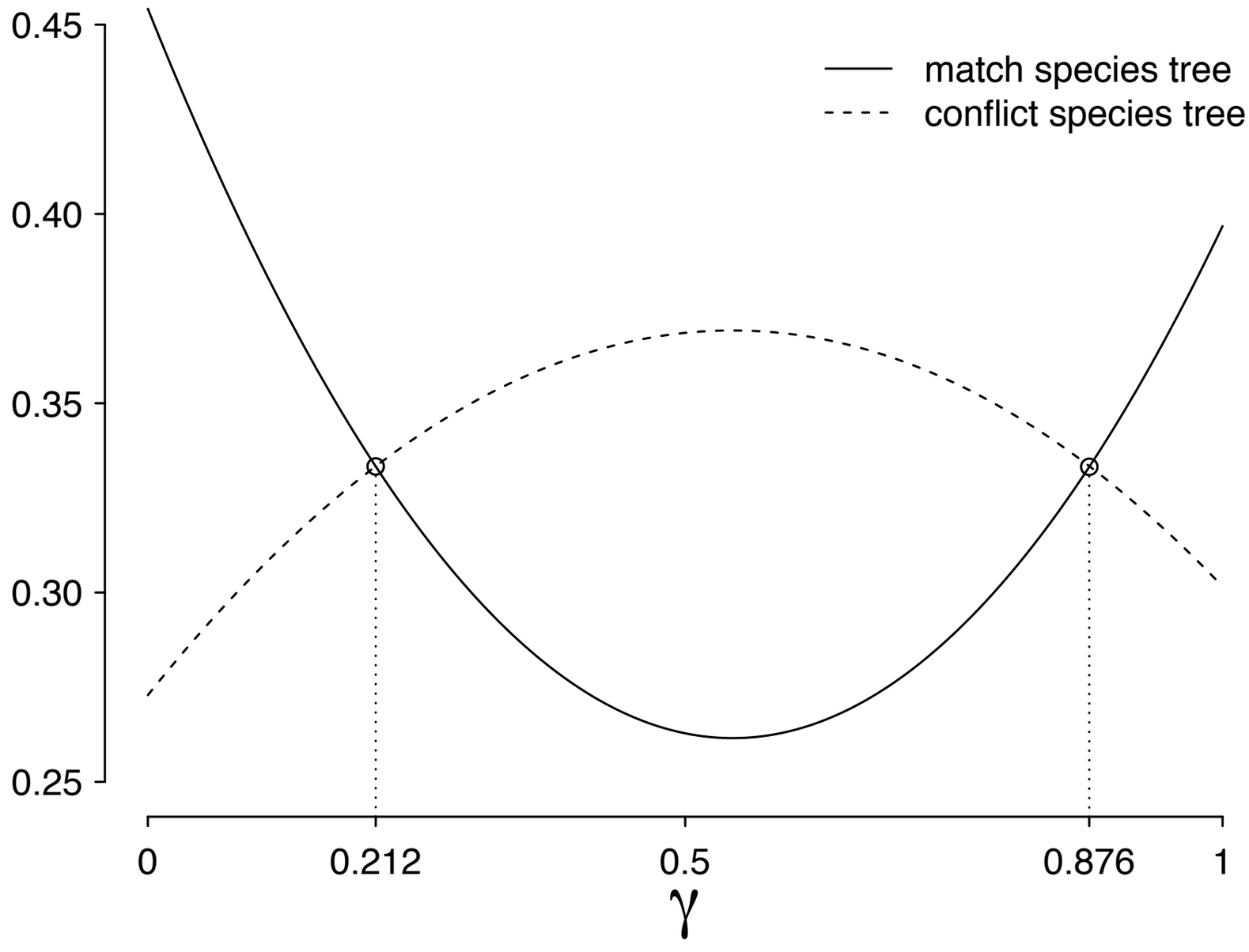
(S.-L., Yang, Ané, 2016, Syst Bio)

ASTRAL (Mirarab et al, 2014)
NJst (Liu&Yu,2011)

Why?

Phylogenetic network

Anomaly zone with gene flow

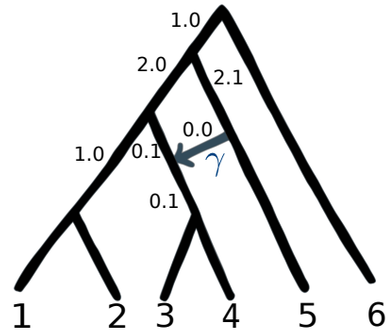


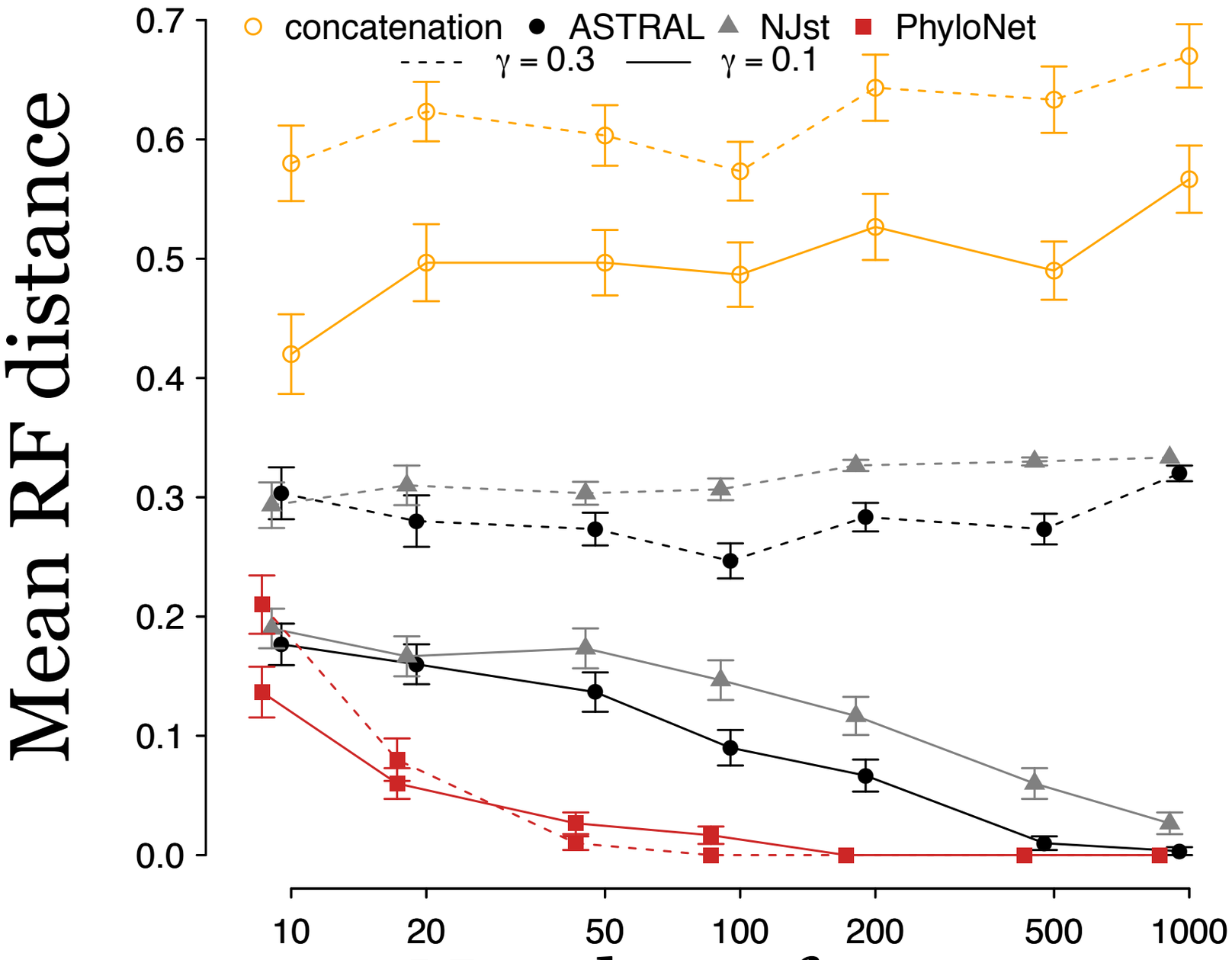
(S.-L., Yang, Ané, 2016, Syst Bio)

Why?

Phylogenetic network

Coalescent tree methods not robust to gene flow





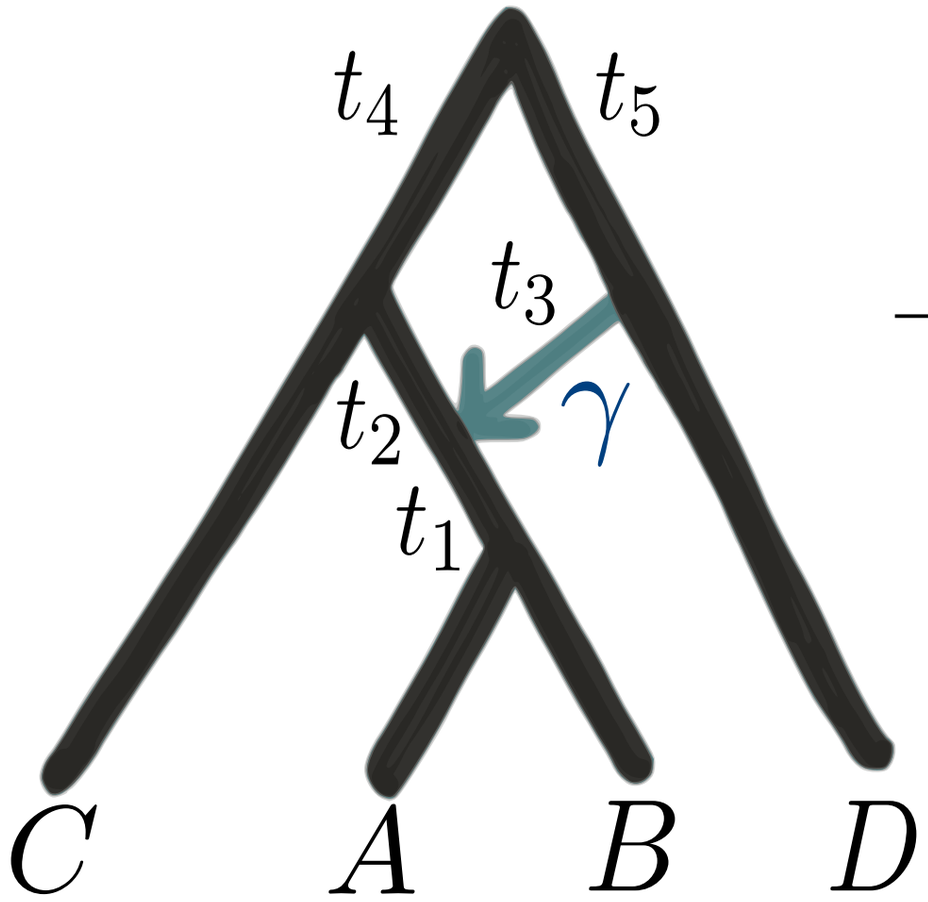
(S.-L., Yang, Ané, 2016, Syst Bio)

ASTRAL (Mirarab et al, 2014)
NJst (Liu&Yu,2011)
PhyloNet (Yu et al 2012, 2014)

Why?

Phylogenetic network

Anomalous unrooted gene trees with gene flow



Frequency among gene trees

Quartet	$\gamma = 0.0$	$\gamma = 0.1$	$\gamma = 0.3$
$AB\|CD$	**0.347**	**0.298**	**0.260**
$CA\|BD$	0.327	0.351	0.370
$CB\|AD$	0.327	0.351	0.370

$$t_1 = t_2 = 0.01, t_3 = t_4 = t_5 = 1$$

- **ILS**: no AUGT on 4 taxa (Degnan, 2013)
- **ILS**+**HGT**: AUGT on 4 taxa (S.-L., Yang, Ané, 2016, Syst Bio)

See also Long & Kubatko (2018) for AUGT under continuous gene flow between sister species

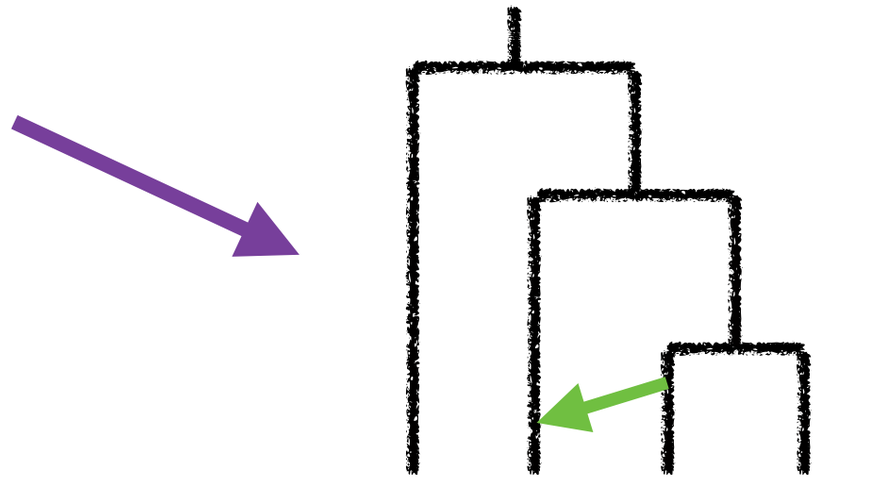
So far…

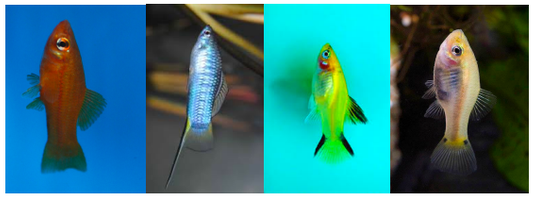
- Networks are good
- Explicit networks are better
- If you ignore gene flow, you can estimate the wrong tree

How?

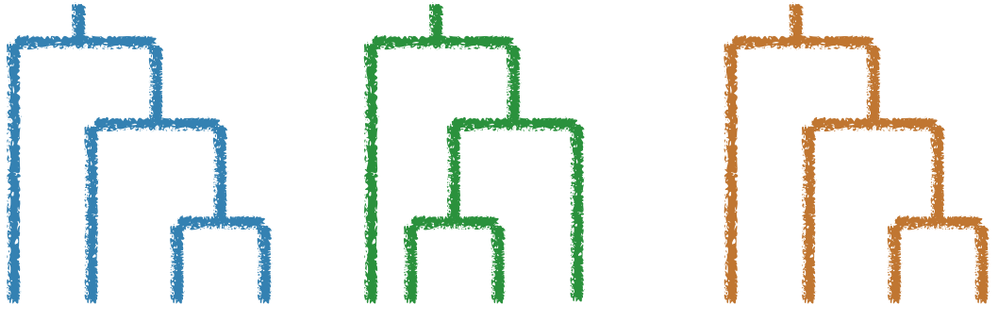
Phylogenetic network

BEAST2
(Zhang et al, 2017)
PhyloNet
(Wen et al, 2016)





MrBayes
(Huelsenbeck, Ronquist, 2001)
RAxML
(Stamatakis, 2014)
PhyML
(Guindon et al, 2010)



SNaQ
(S.-L., Ane, 2016)
PhyloNet
(Yu et al, 2014)

Multispecies coalescent on a network

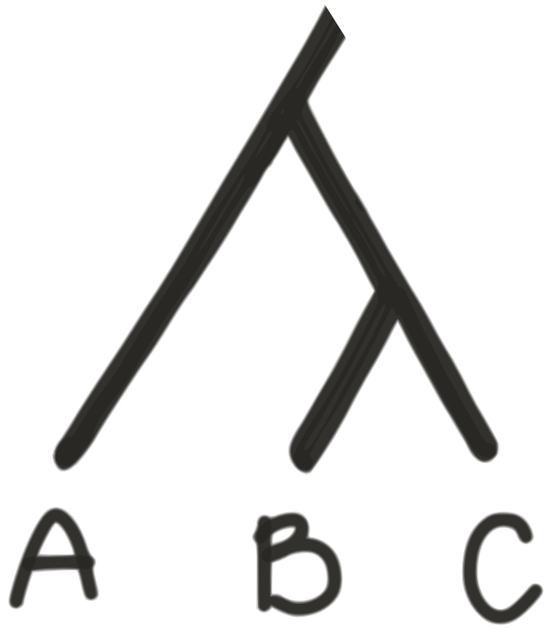
(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)

Multispecies coalescent on a network

(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)

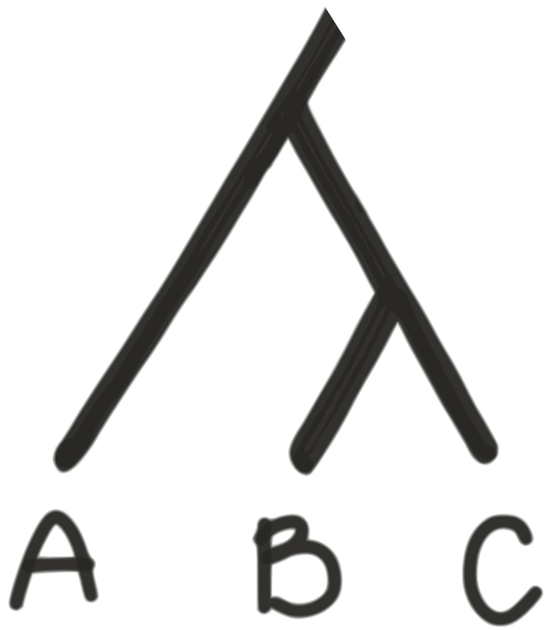
Multispecies coalescent on a network

$$P\left(\quad \middle| \quad \right)$$



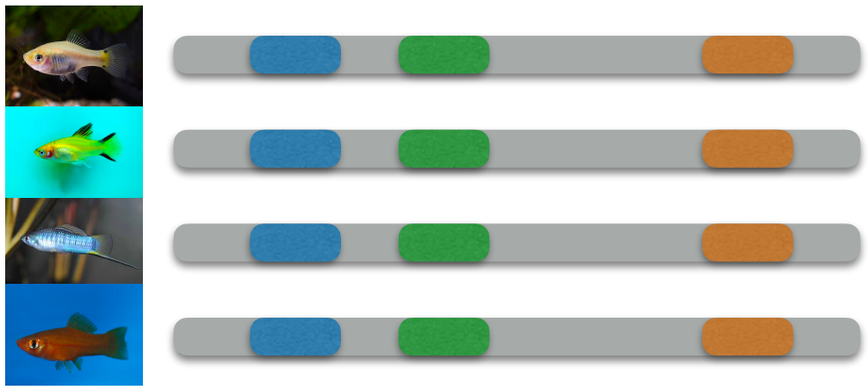
(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)

Multispecies coalescent on a network

$P($  $|$  $)$

$$p_{BC|AD}(t, t_2, \gamma) = (1-\gamma)\frac{1}{3}e^{-t} + \gamma(1 - \frac{2}{3}e^{-t_2})$$

(Meng, Kubatko, 2009)
(Yu, Degnan, Nakhleh, 2012)



$$P(N, G, \theta | D) \propto \pi(N)\pi(\theta) \prod_{i=1}^{L} P(D_i|G_i)P(G_i|N)$$

network → N; gene trees → G; L loci → D

Prior Network → $\pi(N)$

Substitution Model → $P(D_i|G_i)$

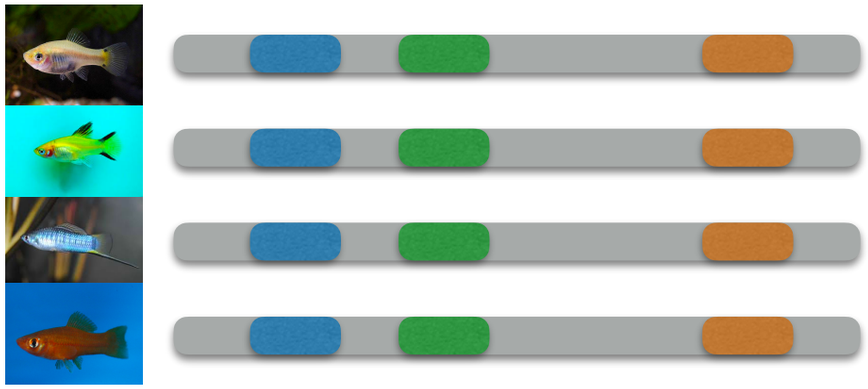
Multispecies Coalescent → $P(G_i|N)$

Birth-hybridization process

reticulations, cycle diameter

BEAST2
(Zhang et al, 2017)

PhyloNet Bayesian
(Wen et al, 2016)



$$P(N, G, \theta|D) \propto \pi(N)\pi(\theta) \prod_{i=1}^{L} P(D_i|G_i)P(G_i|N)$$

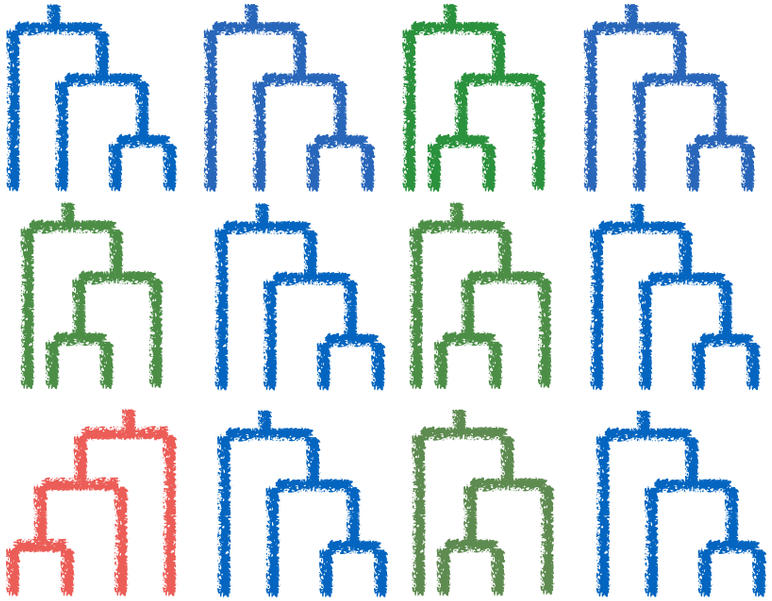
network · gene trees · L loci · **Prior** · **Network** · **Substitution Model** · **Multispecies Coalescent**

Birth-hybridization process

reticulations, cycle diameter

BEAST2
(Zhang et al, 2017)

PhyloNet Bayesian
(Wen et al, 2016)

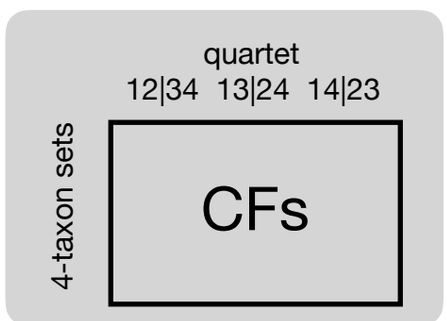


$$P(N, \theta|G) \propto \pi(N)\pi(\theta) \prod_{i=1}^{L} P(G_i|N, \theta)$$

PhyloNet Bayesian
(Wen et al, 2016)

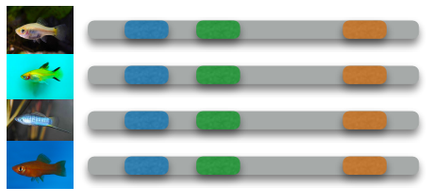
$$L(N, \theta) = \prod_{i=1}^{L} P(G_i|N, \theta)$$

PhyloNet Likelihood
(Yu et al, 2014)



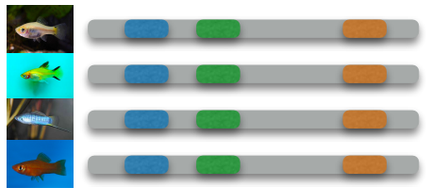
$$\tilde{L}(N, \theta) \propto \prod_{q} L(q|N, \theta)$$

SNaQ
(S.-L., Ane, 2016)



BEAST2
(Zhang et al, 2017)

Birth-hybridization process

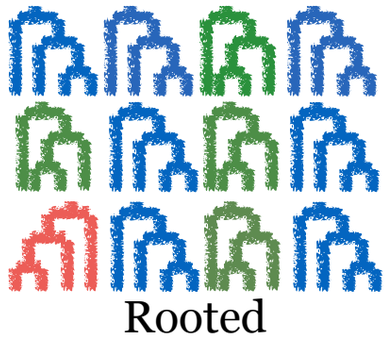


PhyloNet Bayesian
(Wen et al, 2016)

reticulations, cycle diameter

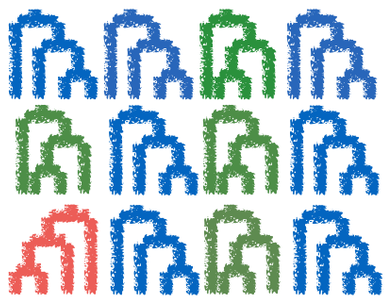
MCMC: Network moves, mixing

Most accurate, not scalable



Rooted

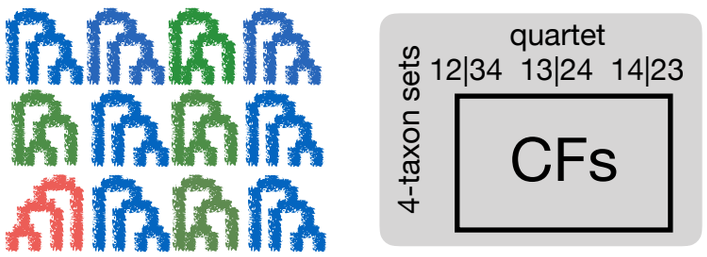
PhyloNet Bayesian
(Wen et al, 2016)



Rooted

PhyloNet Likelihood
(Yu et al, 2014)

Heuristic search: Network moves



Unrooted

SNaQ
(S.-L., Ane, 2016)

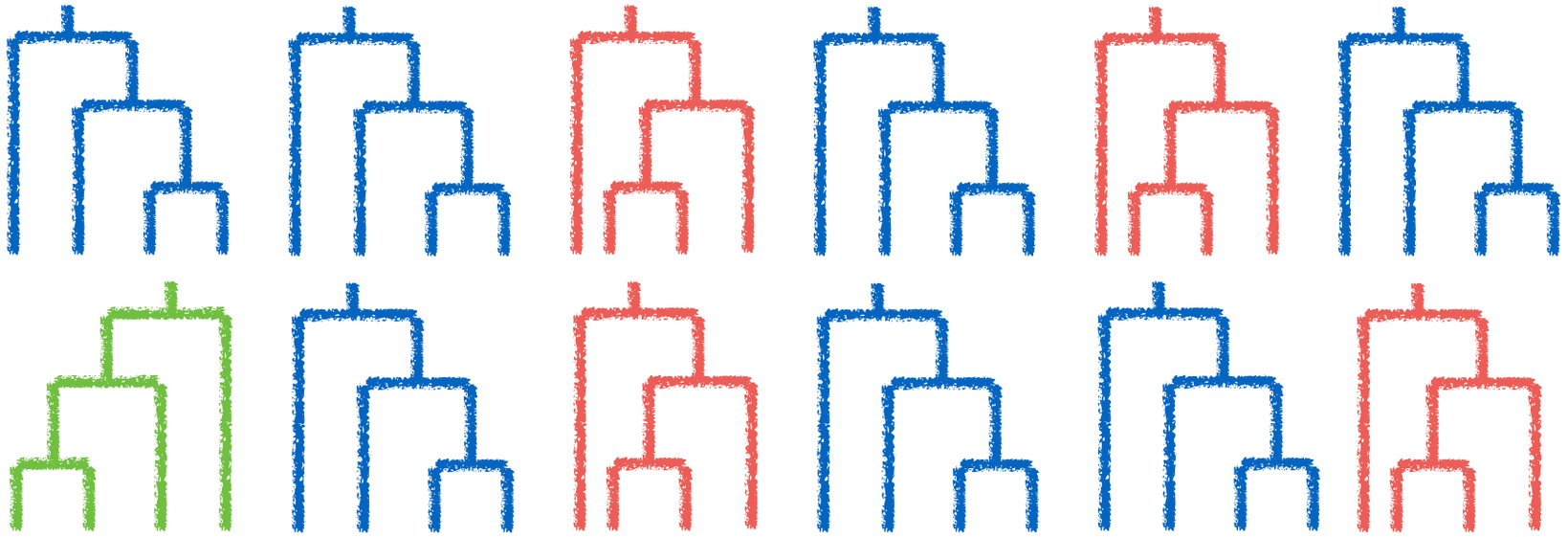
Level-1 networks

More scalable, Robust

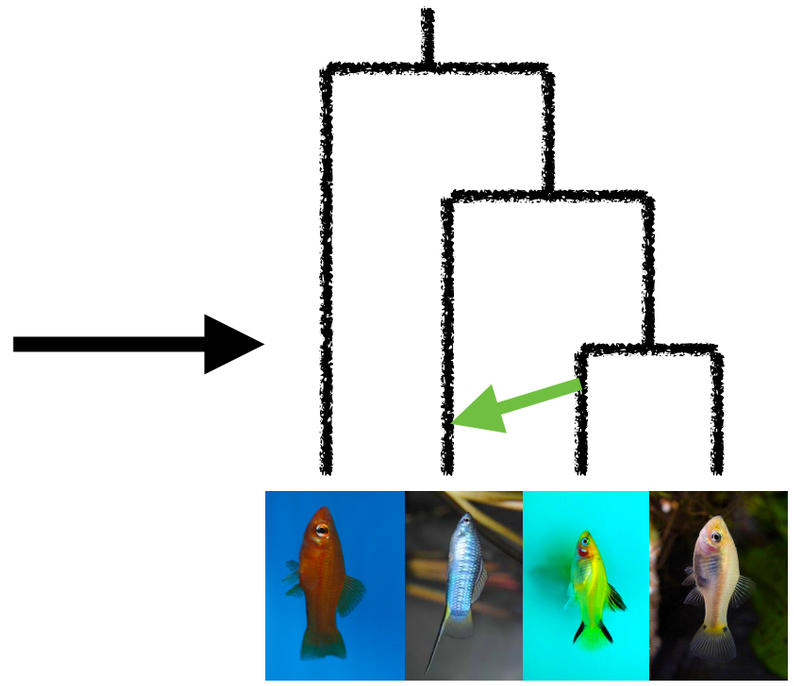
STEM-hy	gene trees rooted, BL	likelihood	hybridization b/w sister lineages
PhyloNet `InferNetwork_ML`	gene trees rooted	likelihood	
PhyloNet `InferNetwork_MPL`	gene trees rooted	triplet likelihood	
PhyloNetworks `SNaQ`	gene trees or quartet CFs	quartet likelihood	level-1 network
PhyloNet `MCMC_GT`	gene trees rooted	Bayesian	compound prior
PhyloNet `MCMC_SEQ`	alignments	Bayesian	compound prior no rate variation
BEAST2 `SpeciesNetwork`	alignments	Bayesian	birth-hyb prior
PhyloNet `MLE_BiMarkers`	biallelic sites	likelihood	compound prior
PhyloNet `MCMC_BiMarkers`	biallelic sites	Bayesian	compound prior
HyDe	sites	invariants	4 taxa, 1 hyb.

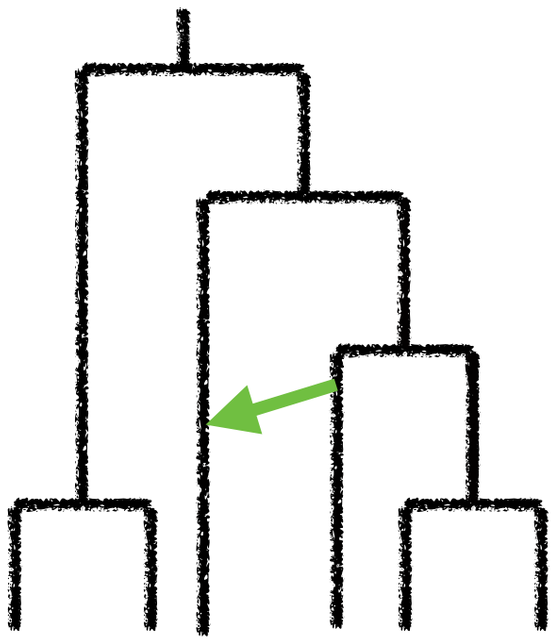
Slide by Cecile Ane

Maximum **pseudo**likelihood



Data





Quartet-based inference

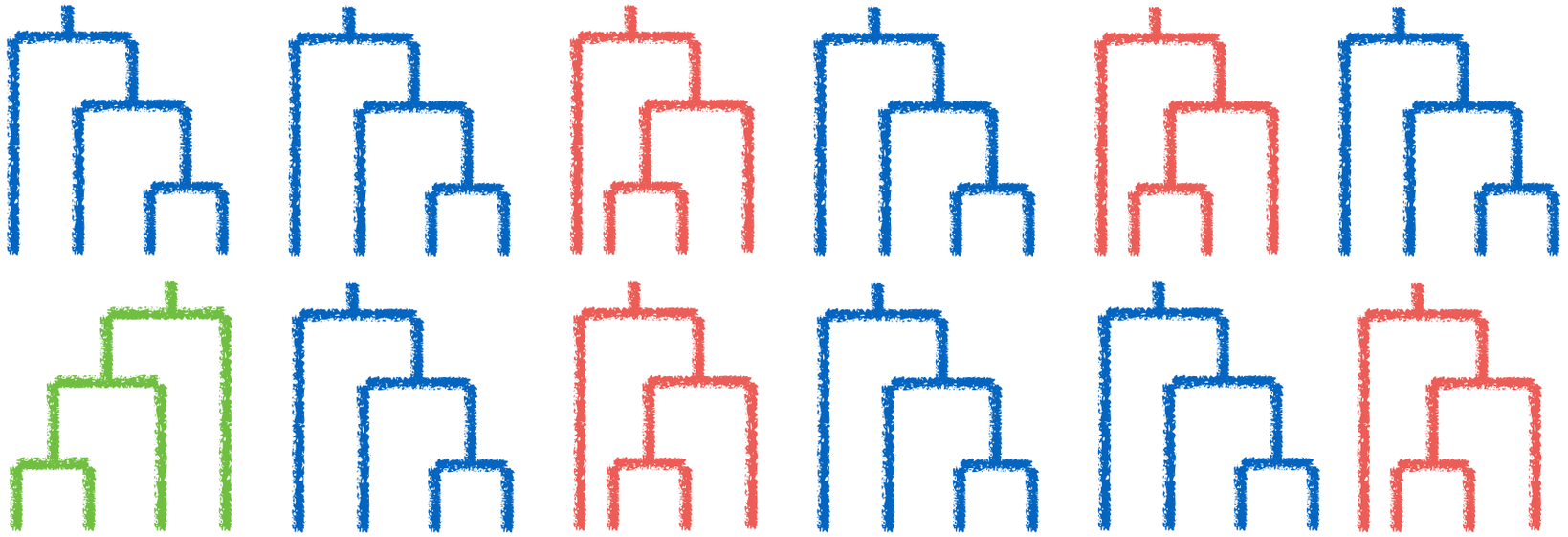
$$\tilde{L}(network) = \prod L(quartet)$$

(S-L, Ané, 2016, PLoS Genetics)

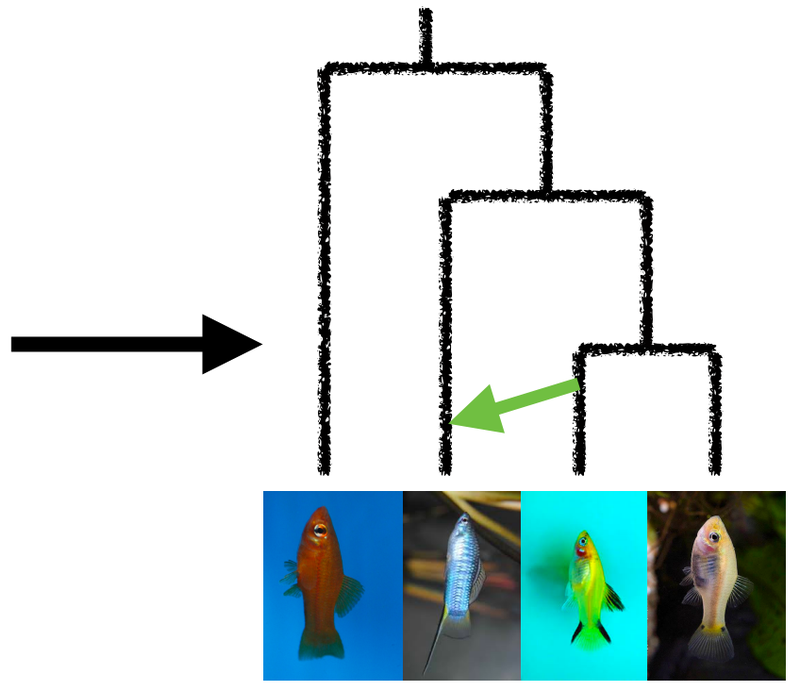
`www.github.com/CRSL4/PhyloNetworks`

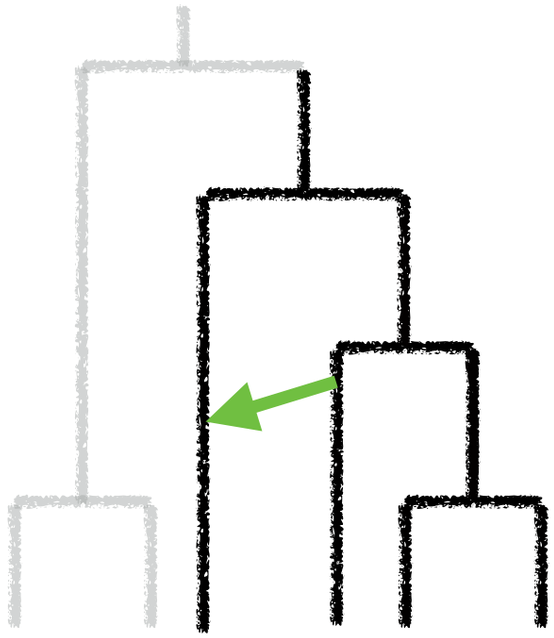
https://solislemuslab.github.io/ @solislemuslab crsl4

Maximum **pseudo**likelihood



Data





Quartet-based inference

$$\tilde{L}(network) = \prod L(quartet)$$

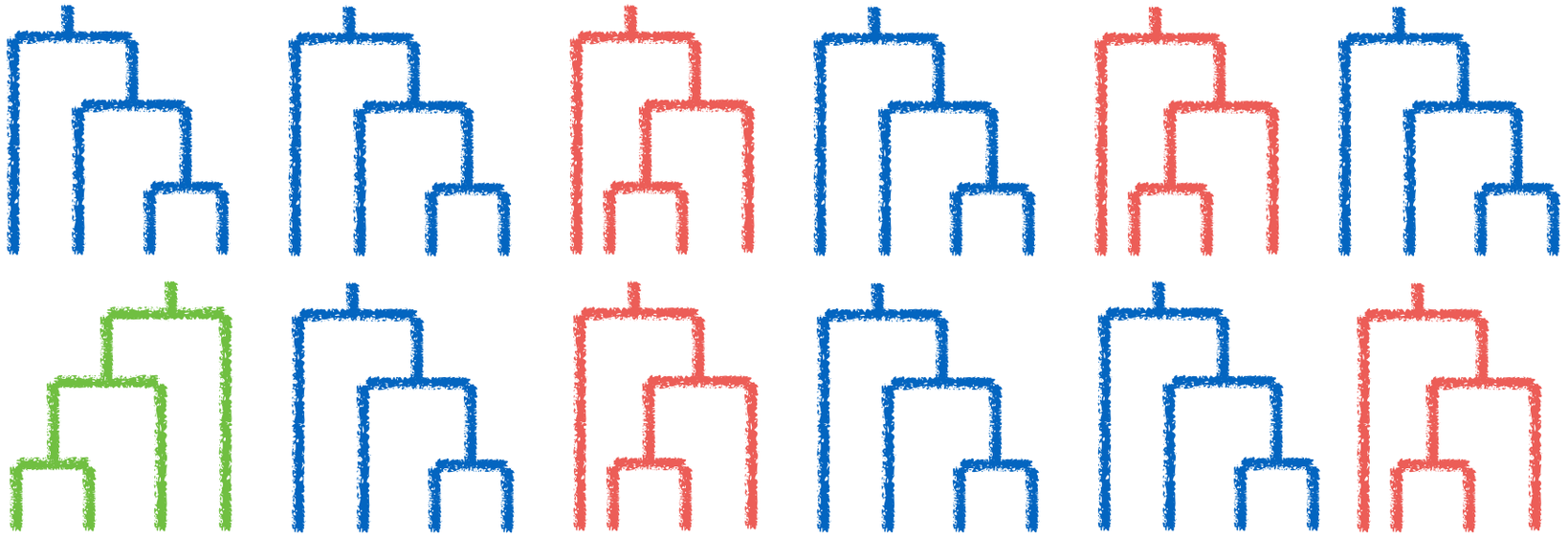
(S-L, Ané, 2016, PLoS Genetics)

`www.github.com/CRSL4/PhyloNetworks`

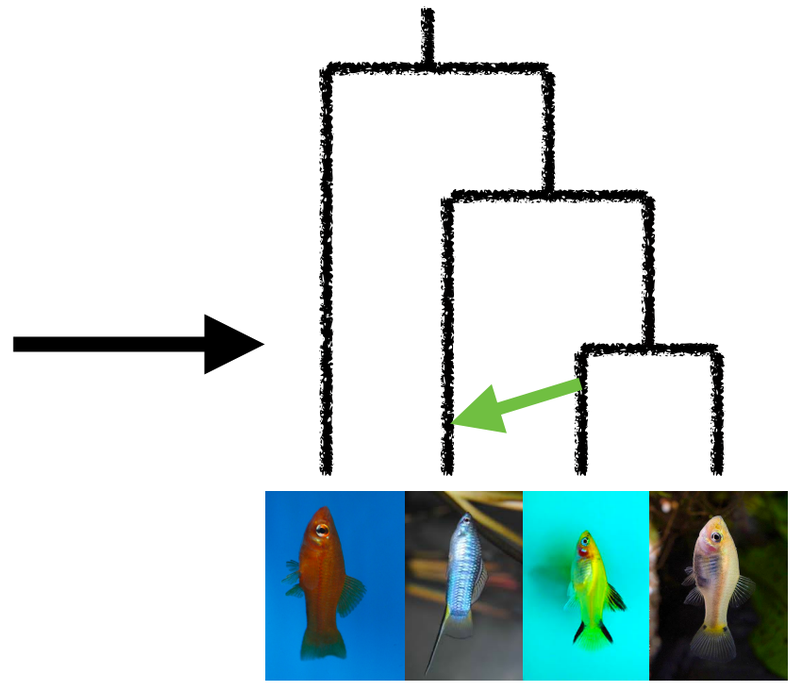
SNAQ julia

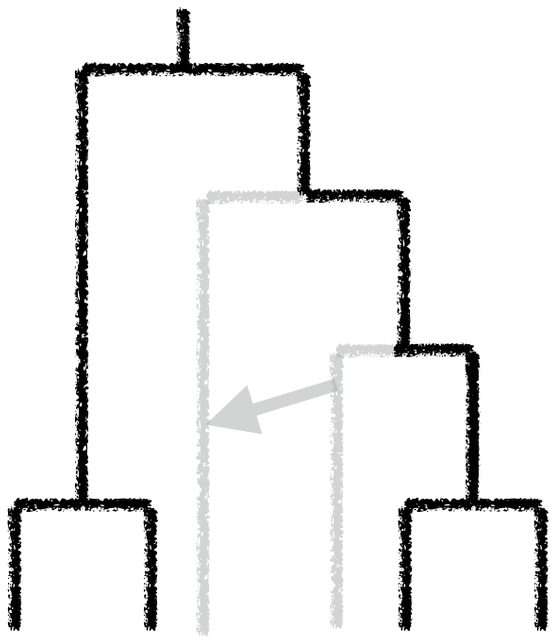
https://solislemuslab.github.io/ @solislemuslab crsl4

Maximum **pseudo**likelihood



Data





Quartet-based inference

$$\tilde{L}(network) = \prod L(quartet)$$

(S-L, Ané, 2016, PLoS Genetics)

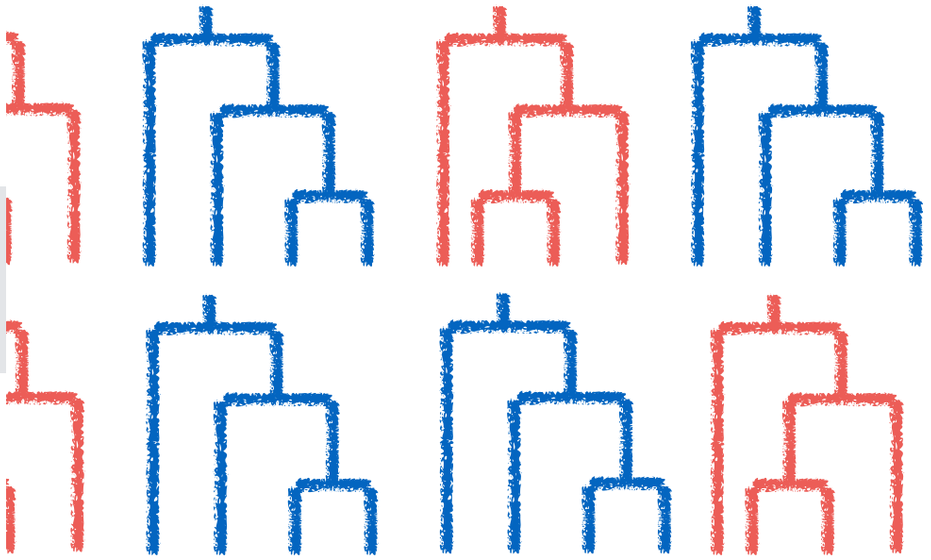
`www.github.com/CRSL4/PhyloNetworks`

SNAQ julia

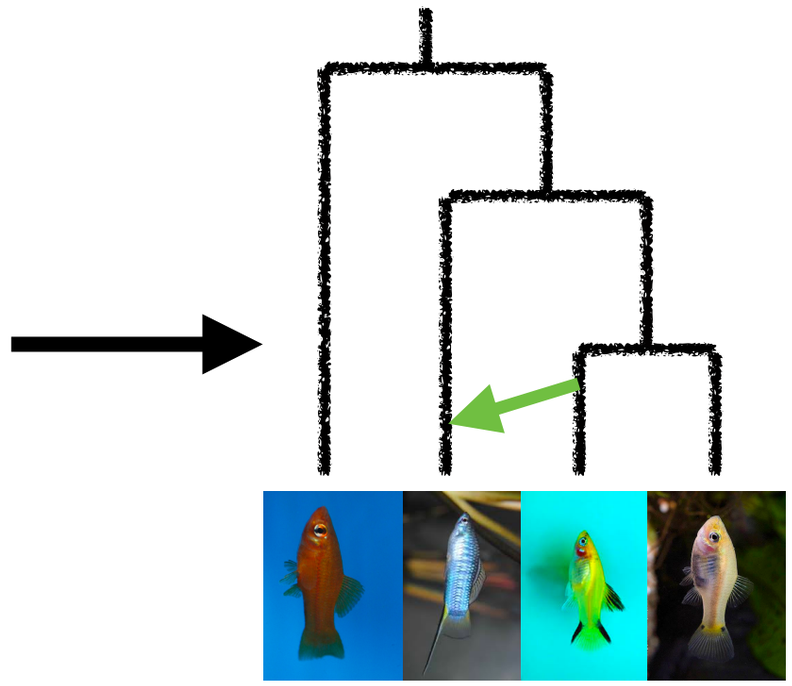
https://solislemuslab.github.io/ @solislemuslab crsl4

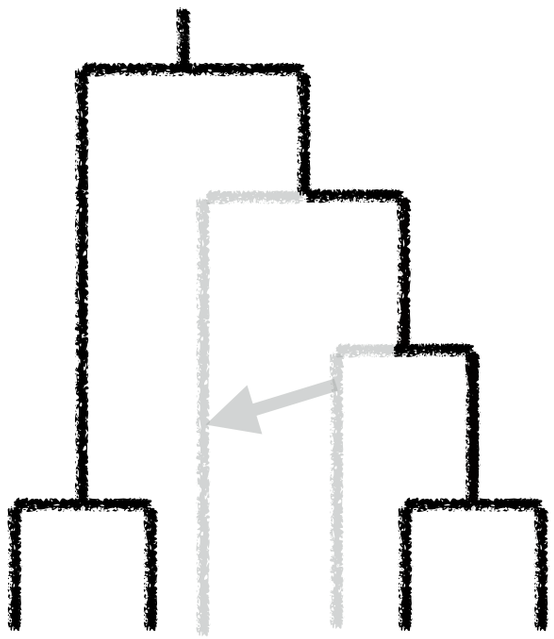
Maximum **pseudo**likelihood

Unrooted gene trees	No rooting error
No branch lengths	No molecular clock assumption
Concordance factors	Account for tree estimation error



Data





Quartet-based inference

$$\tilde{L}(network) = \prod L(quartet)$$

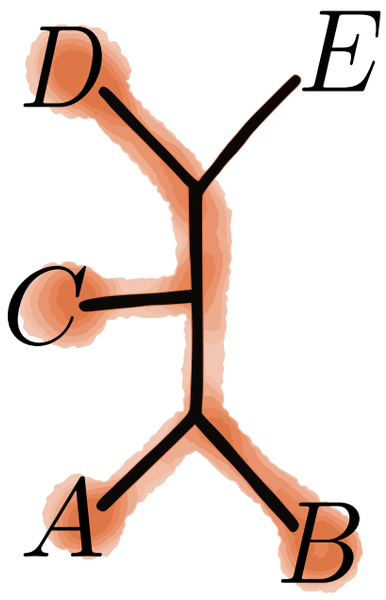
(S-L, Ané, 2016, PLoS Genetics)

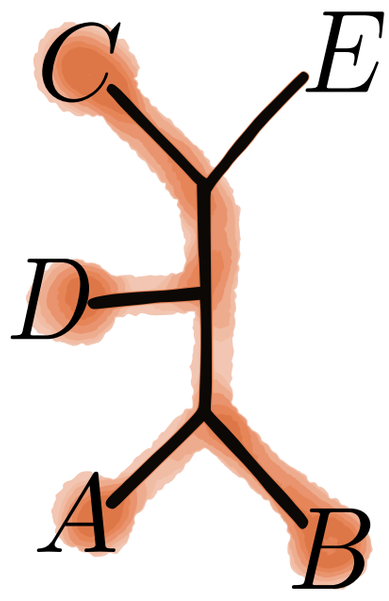
`www.github.com/CRSL4/PhyloNetworks`

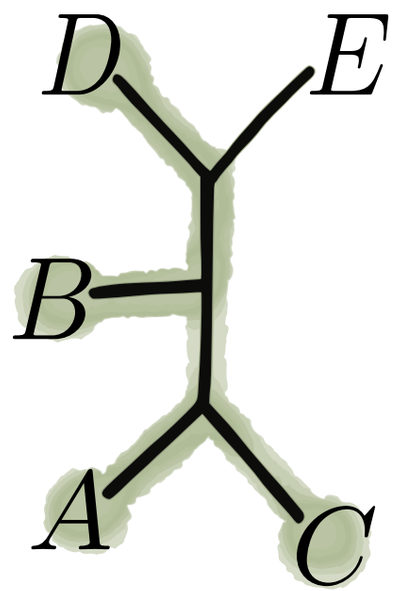
SNAQ julia

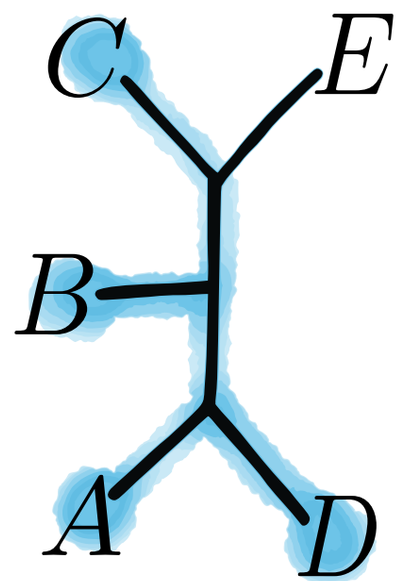
https://solislemuslab.github.io/ @solislemuslab crsl4

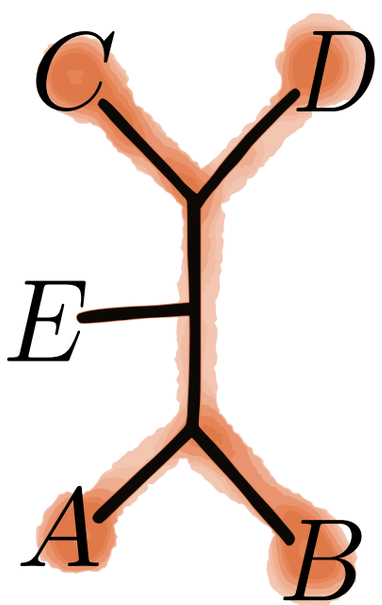
Quartet-based inference



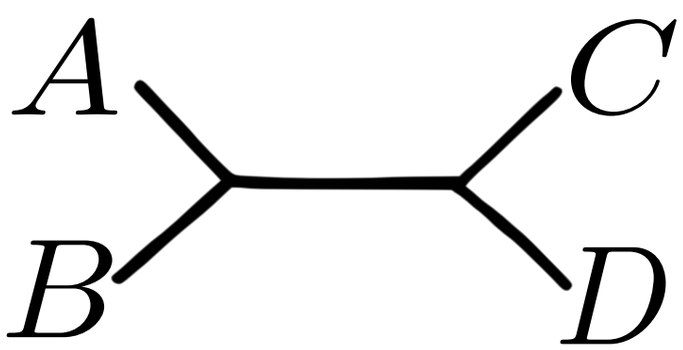




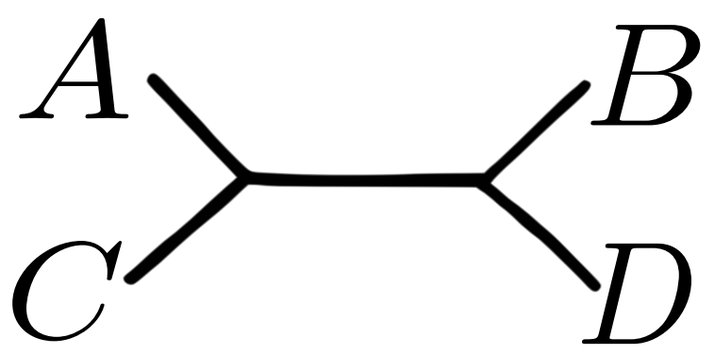




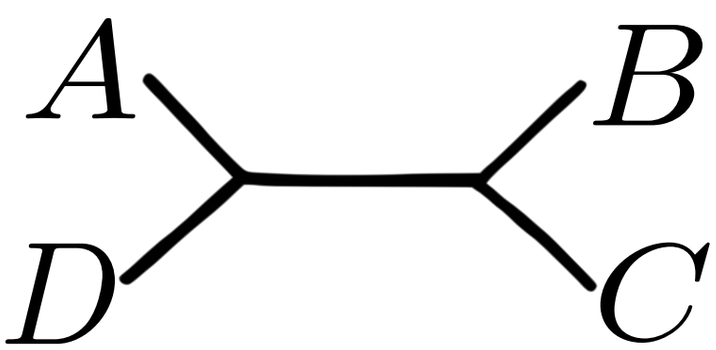
Concordance factors (CF):
% of genes having the quartet in their tree



3/5



1/5



1/5

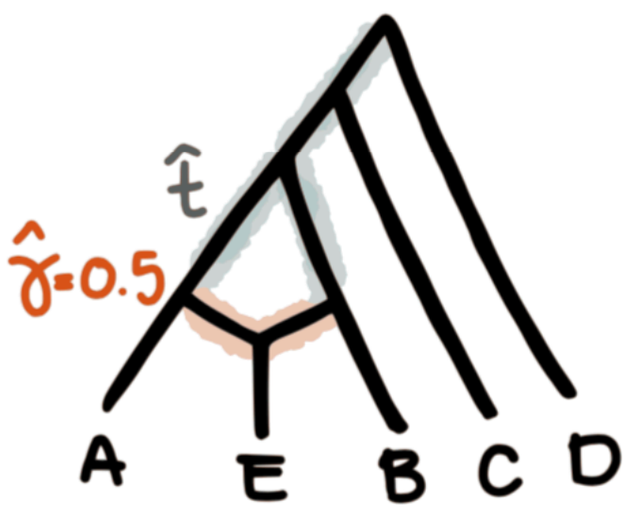
Quartet-based inference

Observed **quartet** CFs:

4 taxon set	CF_1	CF_2	CF_3
A B C D	.80	.10	.10
A B C E	.40	.40	.20
A B D E	.40	.40	.20
A C D E	.84	.08	.08
B C D E	.82	.10	.08

$\longrightarrow$

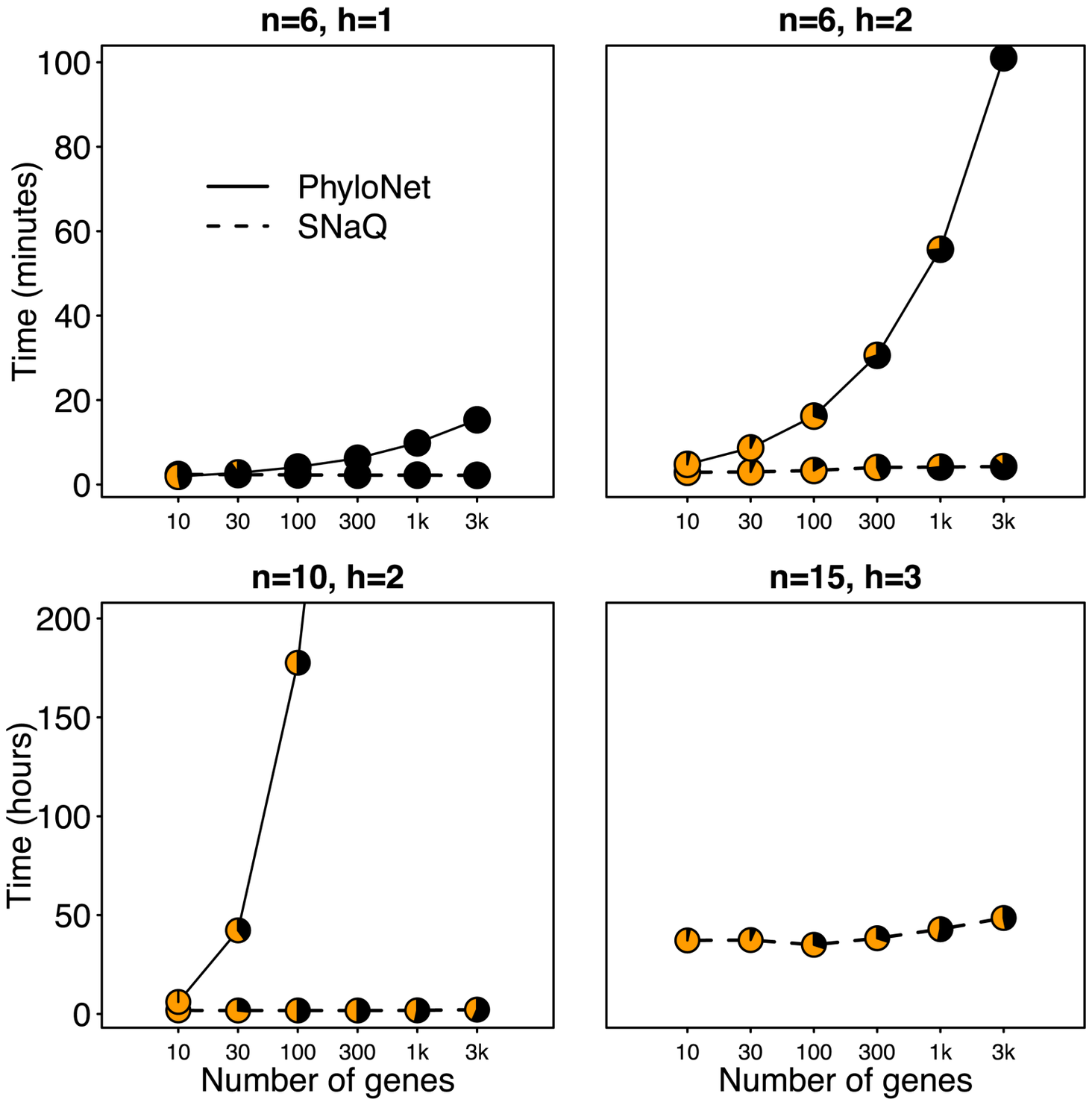
inferred network:



Maximum Pseudo-Likelihood:

$$\log \tilde{L} = \sum_{q \in Q(N)} \mathrm{CF}_{in,1} \log(\mathrm{CF}_{net,1}) + \mathrm{CF}_{in,2} \log(\mathrm{CF}_{net,2}) + \mathrm{CF}_{in,3} \log(\mathrm{CF}_{net,3})$$

Scalability gains

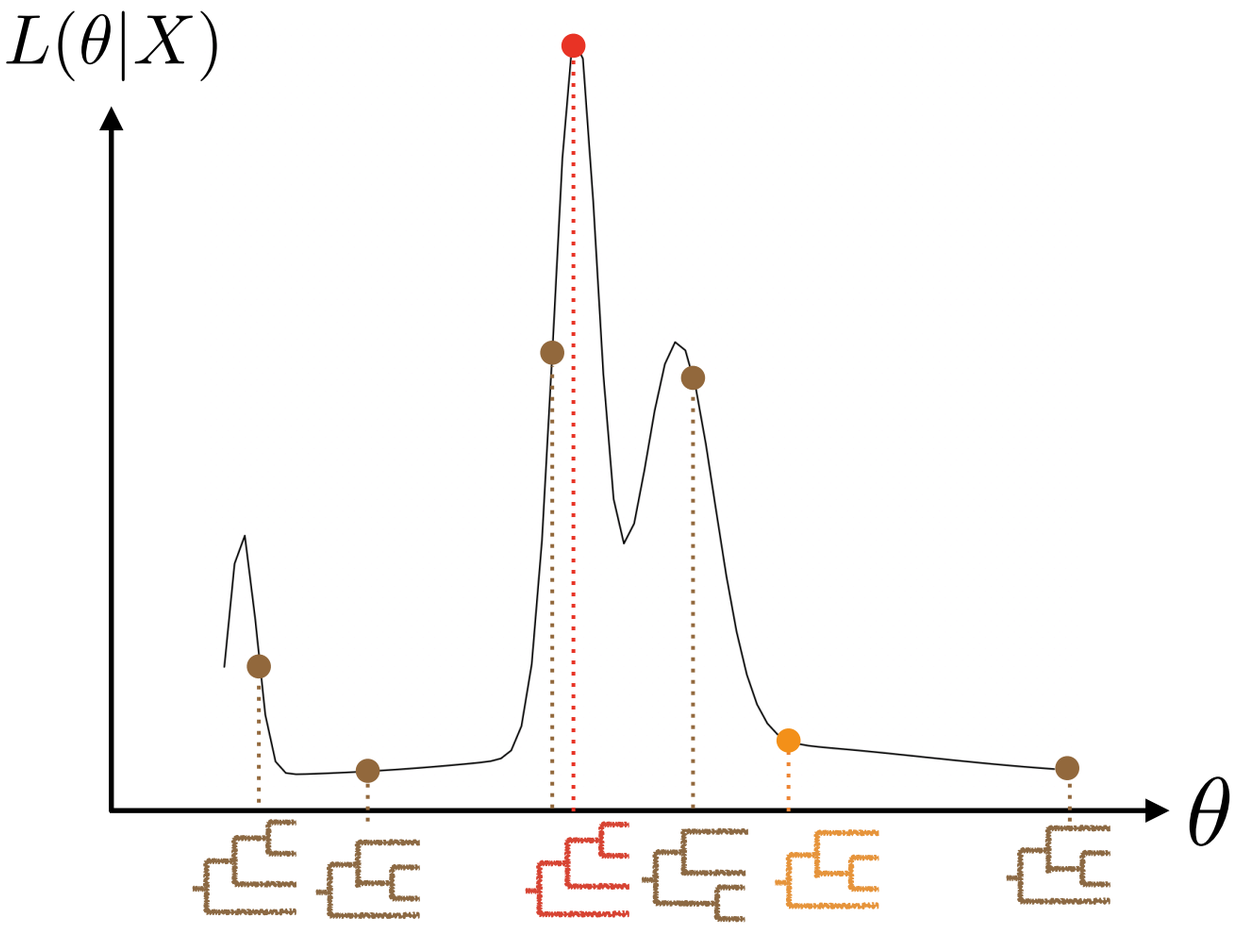


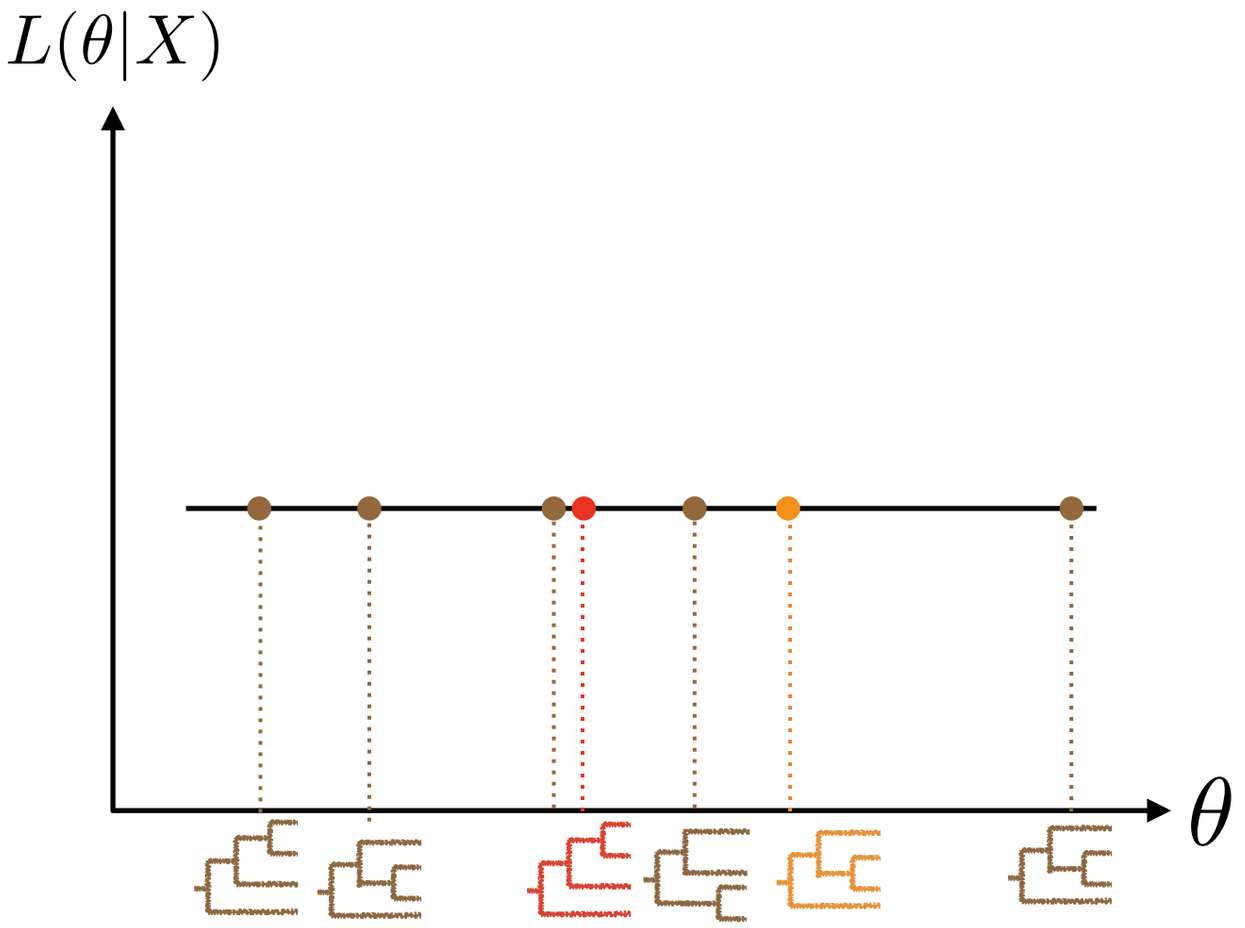
(Solís-Lemus, Ané, 2016, PLoS Genetics)

Challenges

- Network space
- Identifiability
- Network comparison

Identifiability





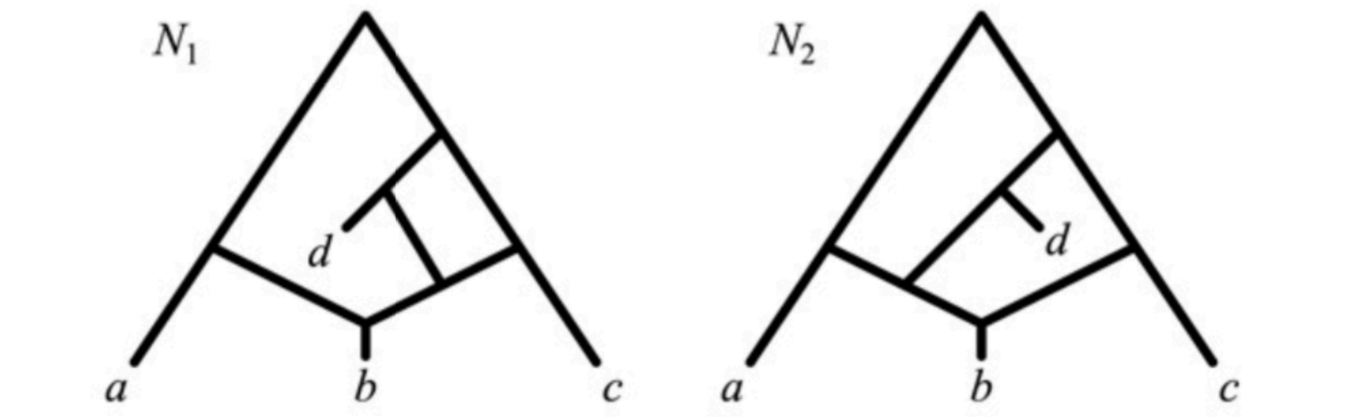
PLOS | COMPUTATIONAL BIOLOGY

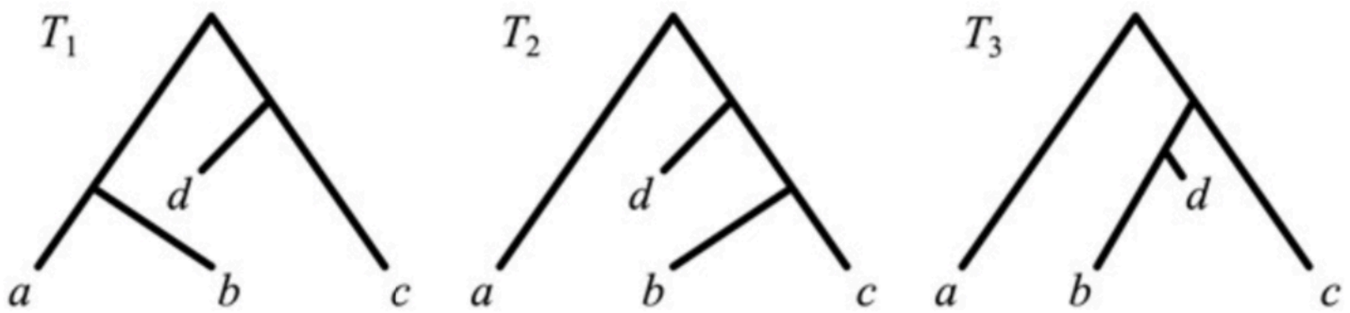
RESEARCH ARTICLE

Reconstructible Phylogenetic Networks: Do Not Distinguish the Indistinguishable

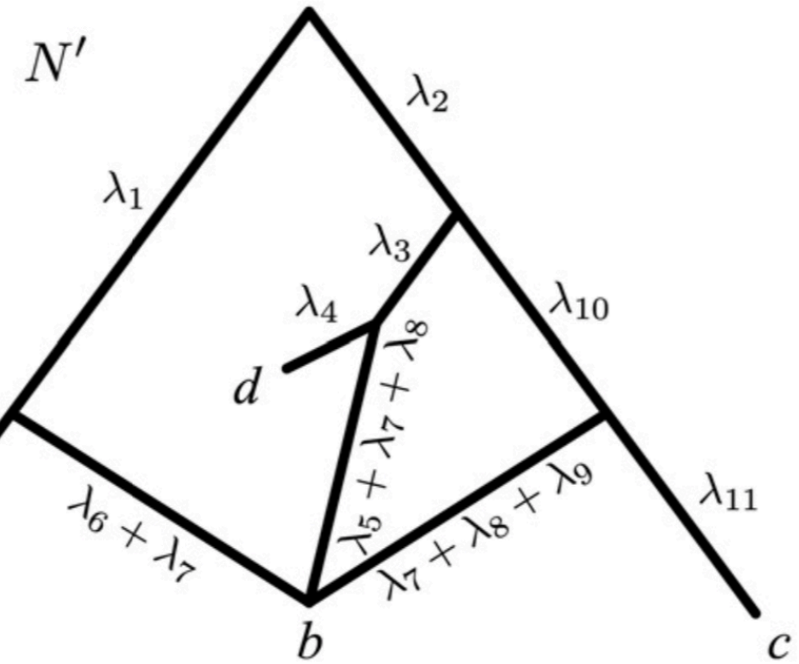
Fabio Pardi[1,3]*, **Celine Scornavacca**[2,3]

1 Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM, UMR 5506) CNRS, Université de Montpellier, France, **2** Institut des Sciences de l'Evolution de Montpellier (ISE-M, UMR 5554) CNRS, IRD, Université de Montpellier, France, **3** Institut de Biologie Computationnelle, Montpellier, France





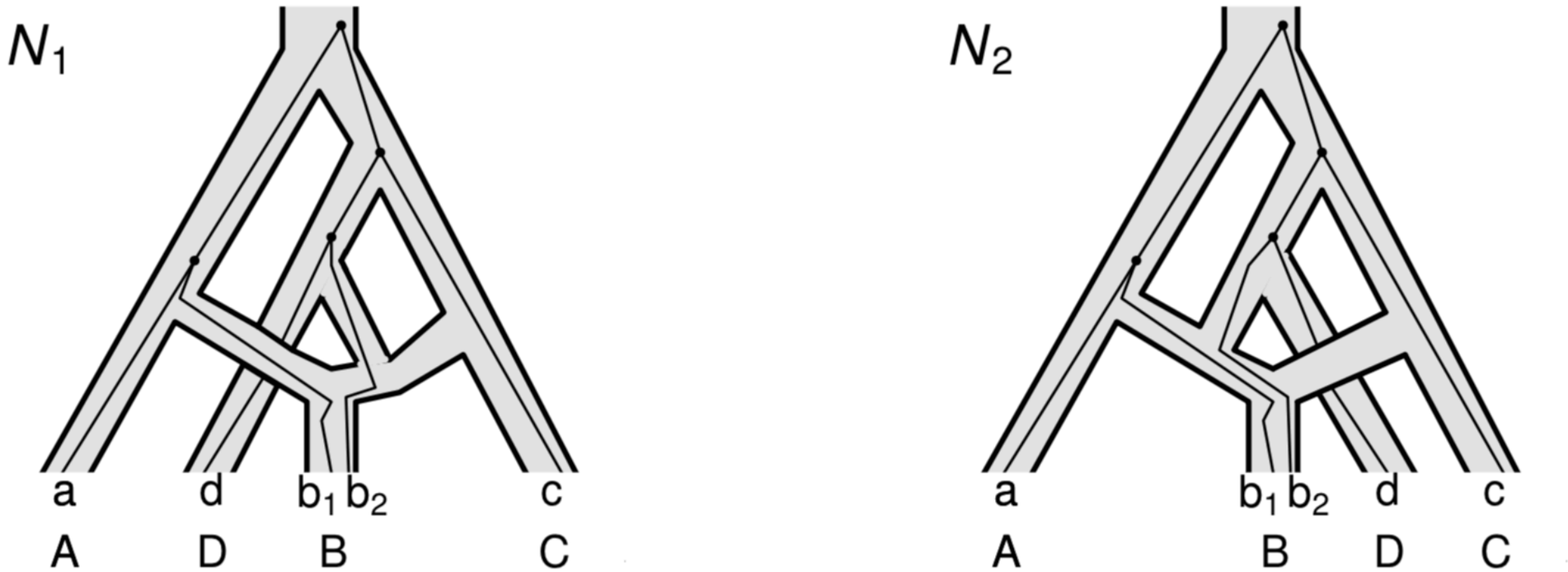
Undistinguishable with the "displayed trees" criterion



Solution: Canonical network ("unzipped")

Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent

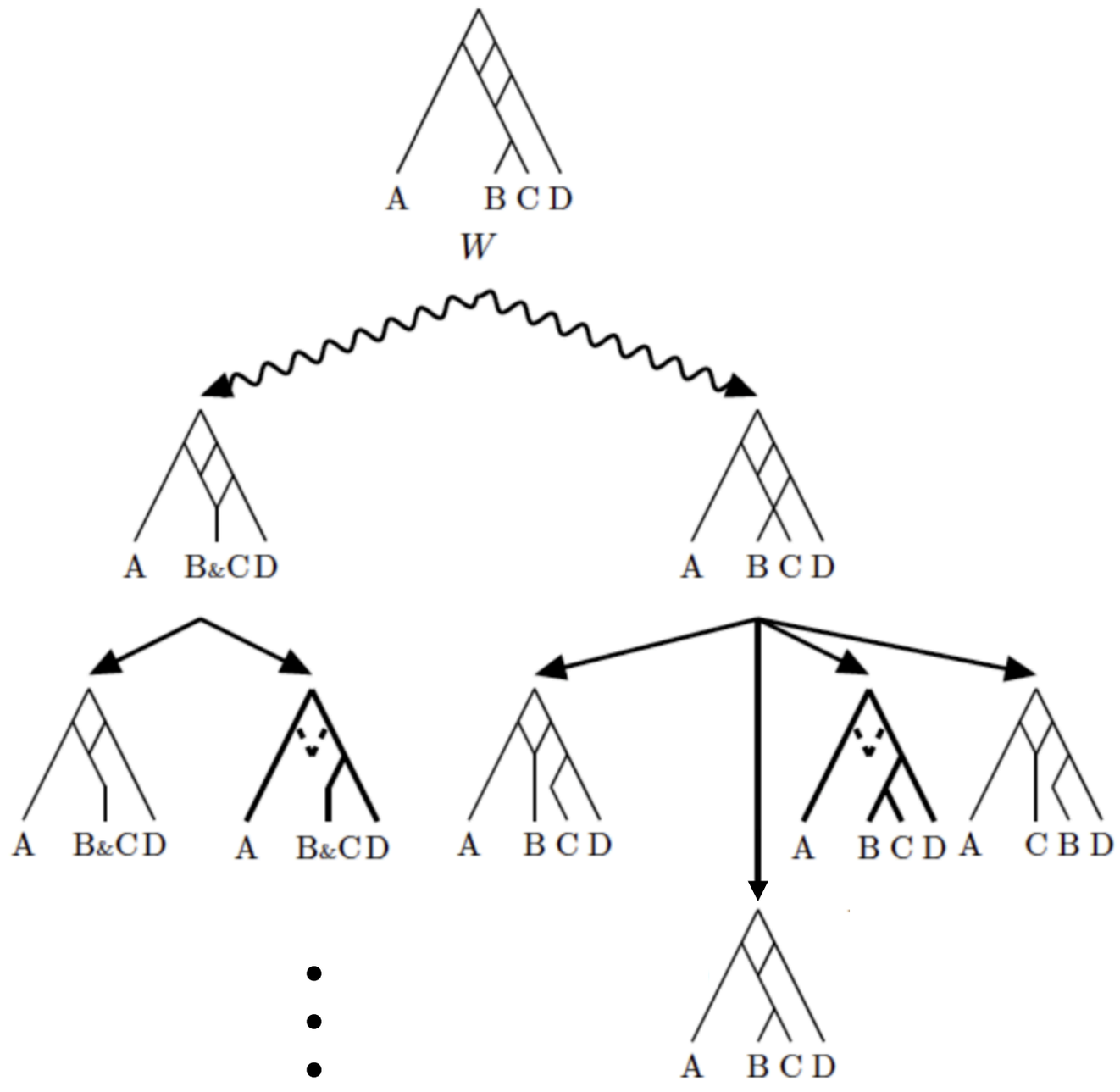
Sha Zhu[1], James H. Degnan[2]



Distinguishable under the MSC

Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent

Sha Zhu[1], James H. Degnan[2]



Decomposing network in **parental** trees

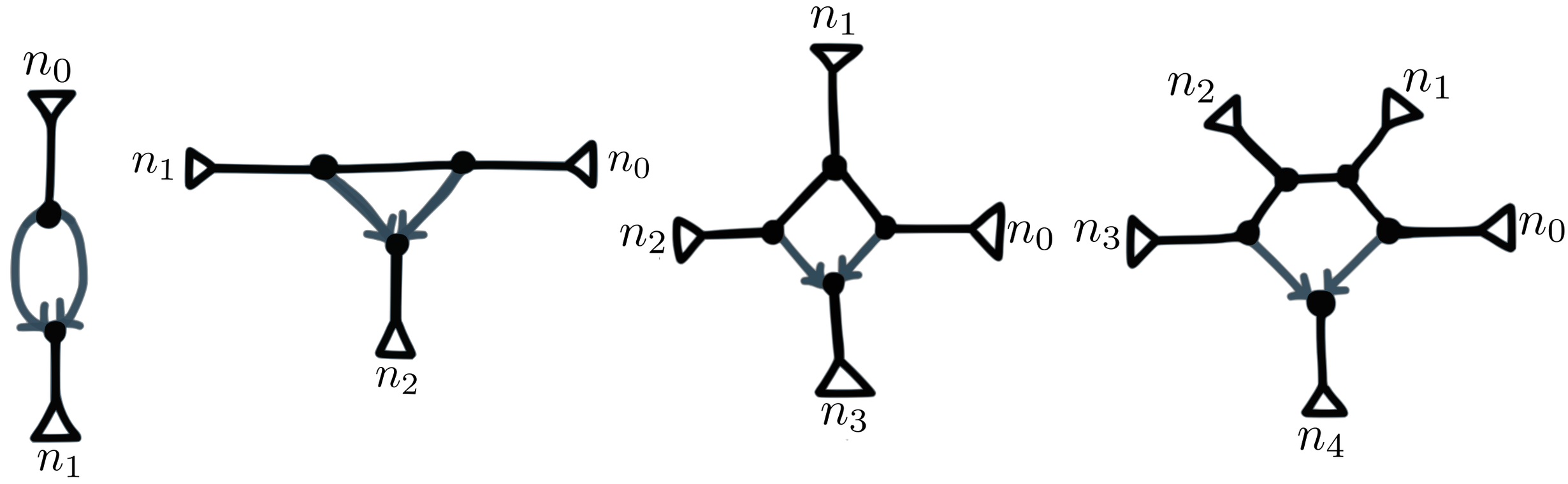
PLOS | GENETICS

RESEARCH ARTICLE

Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting

Claudia Solís-Lemus[1]*, Cécile Ané[1,2]

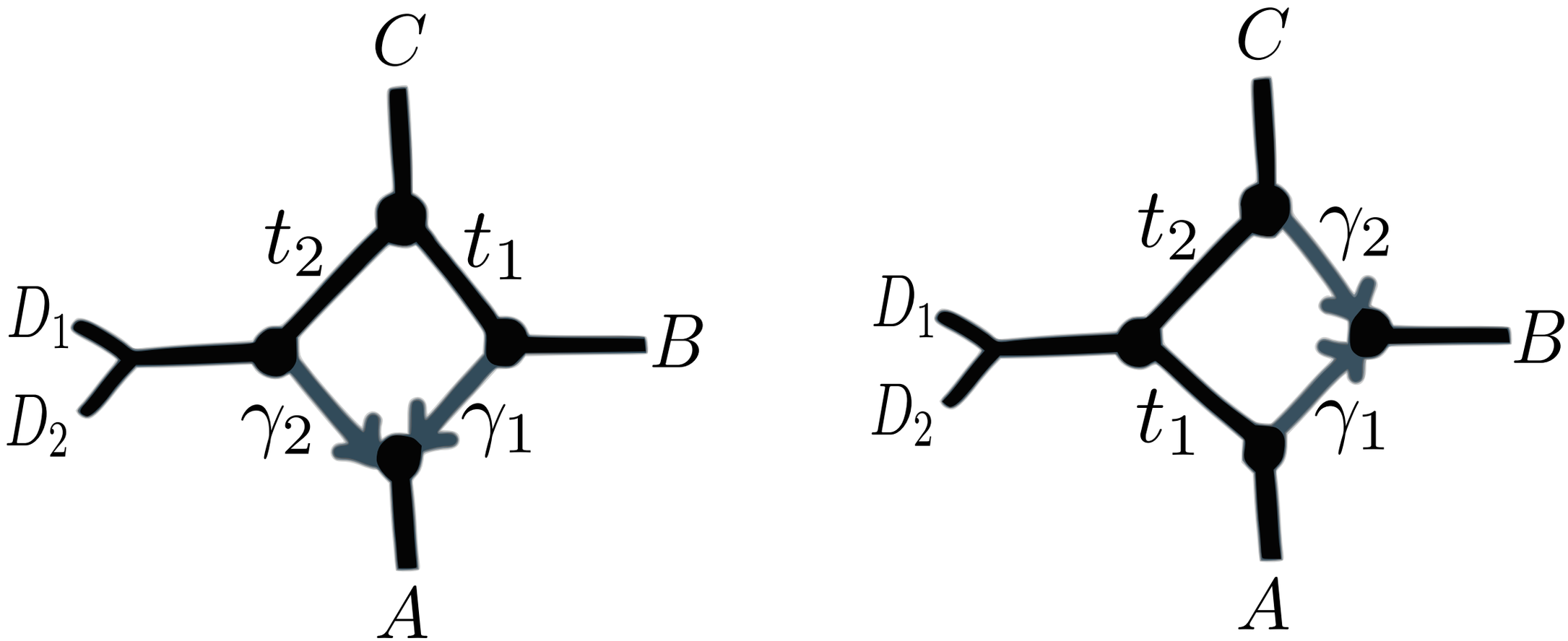
Can we detect the presence of hybridization in level-1 networks?



No **Yes** $(n_i, n_j \geq 2)$ **Yes** $(n_i \geq 2)$ **Yes**

Generic Identifiability $t_i \in (0, \infty), \gamma \in (0, 1)$

In practice: flat pseudolikelihood



(S.-L., Ané, 2016, PLoS Genetics)

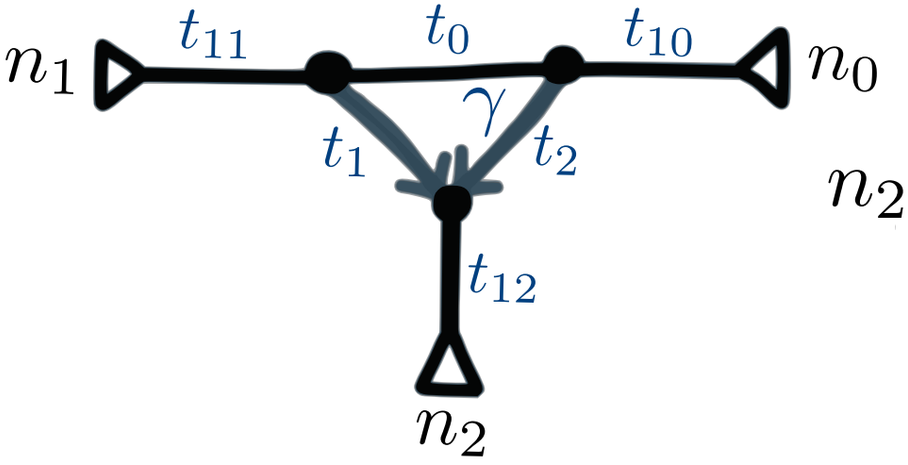
PLOS | GENETICS

RESEARCH ARTICLE

Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting

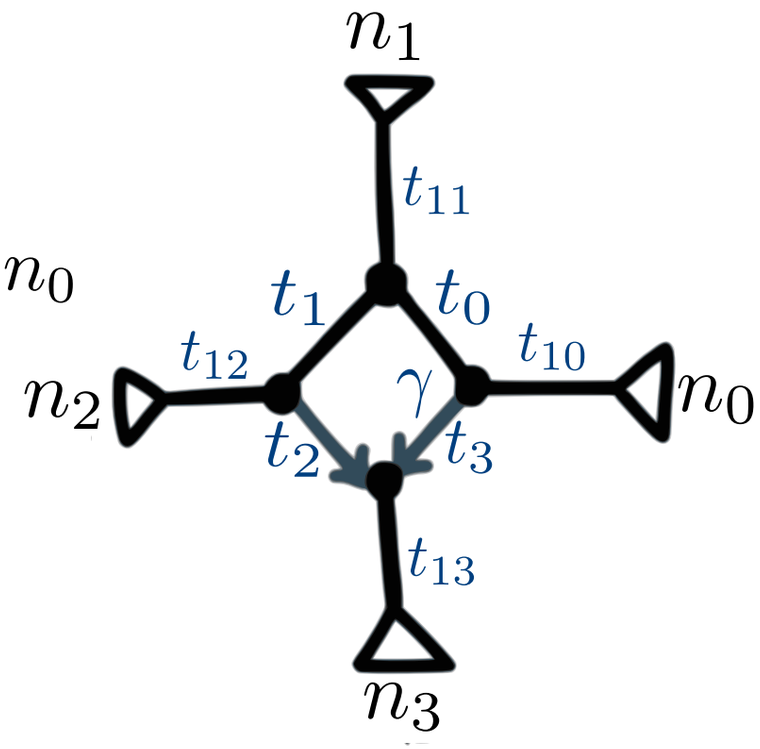
Claudia Solís-Lemus[1]*, Cécile Ané[1,2]

Can we estimate numerical parameters?



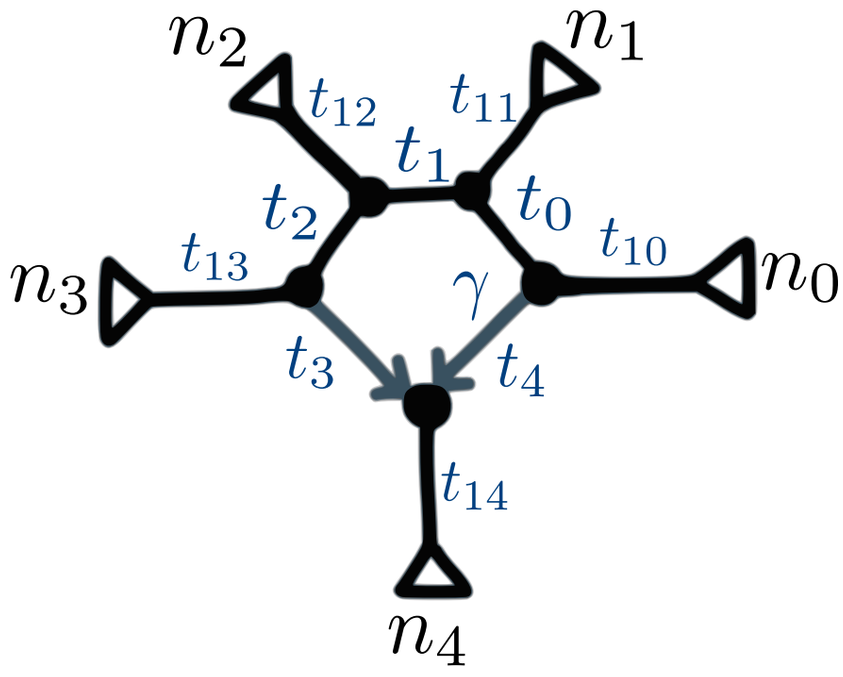
No

Good triangle
$(t_{12} = 0)$



Yes

Good diamond
$(n_0, n_2 \geq 2)$

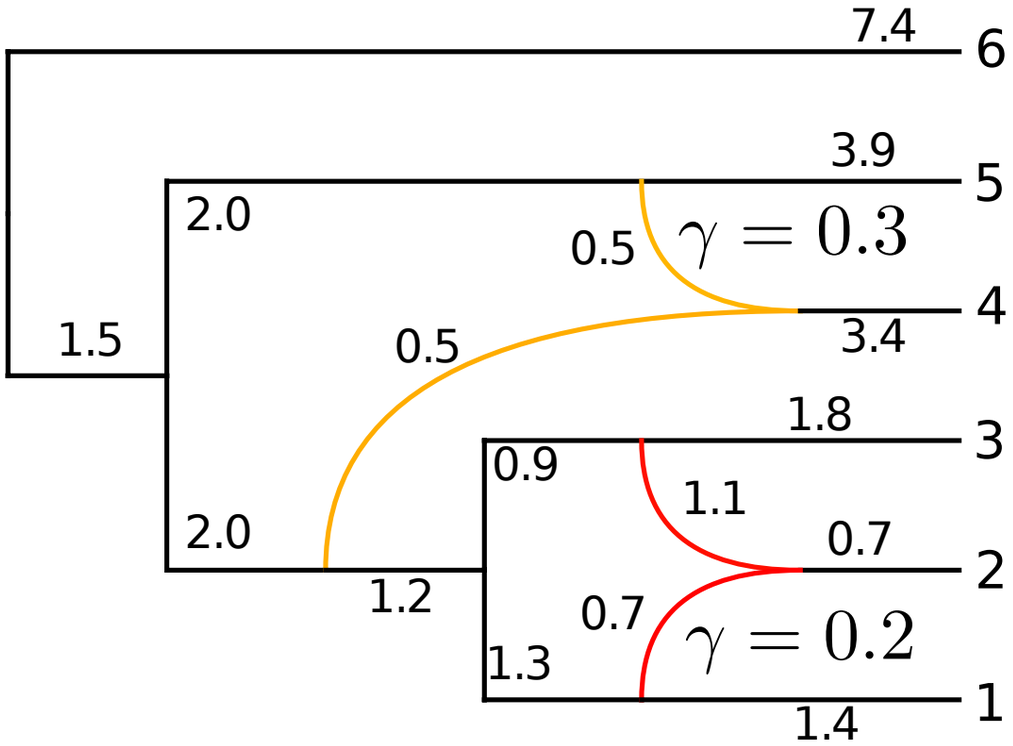


Yes

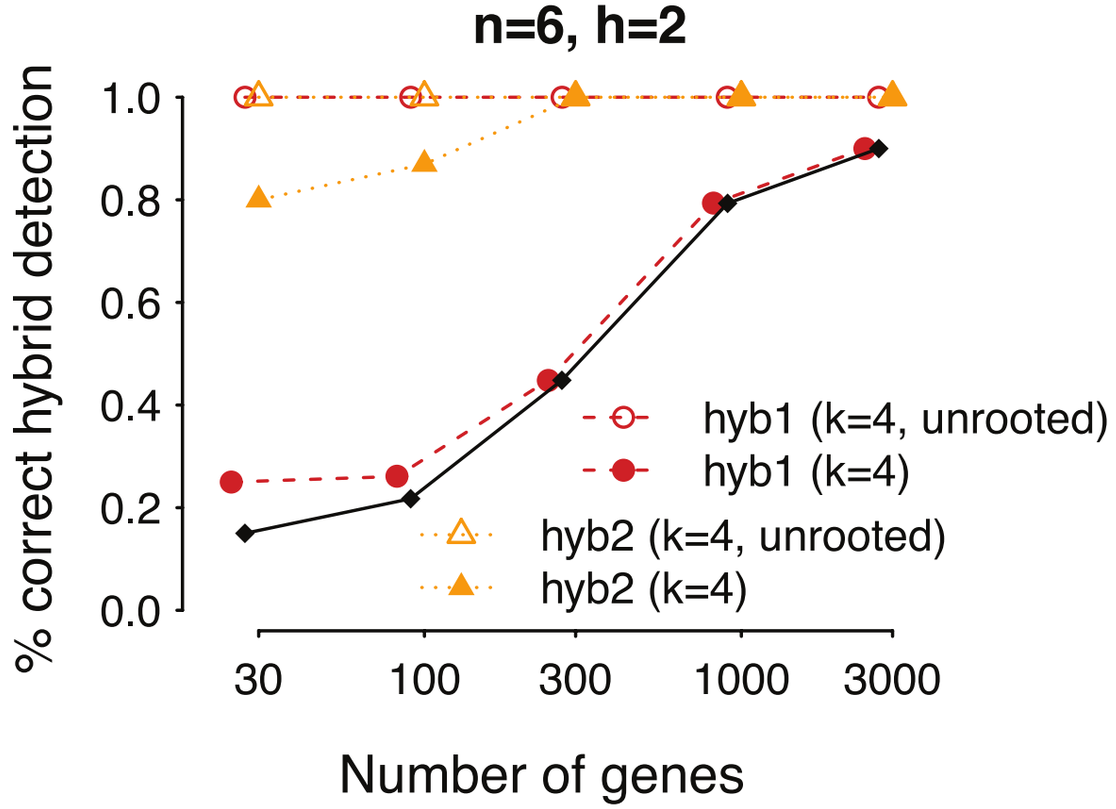
Generic Identifiability $t_i \in (0, \infty), \gamma \in (0, 1)$

Identifiability matters: SNaQ performance

Good diamond



Bad diamond



(S.-L., Ané, 2016, PLoS Genetics)

Challenges

- Network space
- Identifiability

Displayed vs Parental trees
Level-1 semi-directed networks
Hybridizations: case by case
Missing: likelihood, level-k semi-directed

- Network comparison

Challenges

- Network space

K. Huber, V. Moulton, C. Scornavacca,…
Missing: path through tree space, semi-directed

- Identifiability

Displayed vs Parental trees
Level-1 semi-directed networks
Hybridizations: case by case
Missing: likelihood, level-k semi-directed

- Network comparison

Challenges

- Network space

K. Huber, V. Moulton, C. Scornavacca,…
Missing: path through tree space, semi-directed

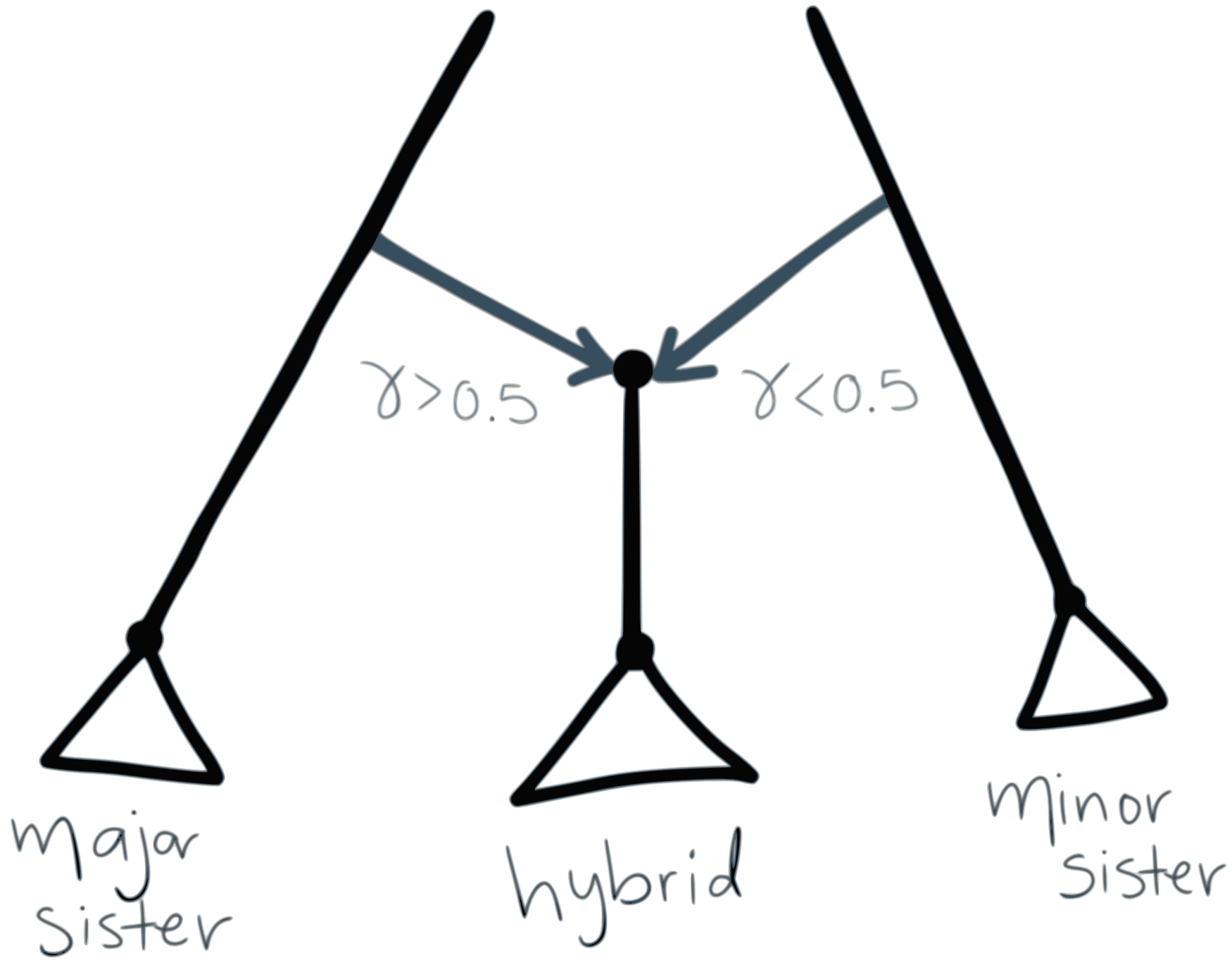
- Identifiability

Displayed vs Parental trees
Level-1 semi-directed networks
Hybridizations: case by case
Missing: likelihood, level-k semi-directed

- Network comparison

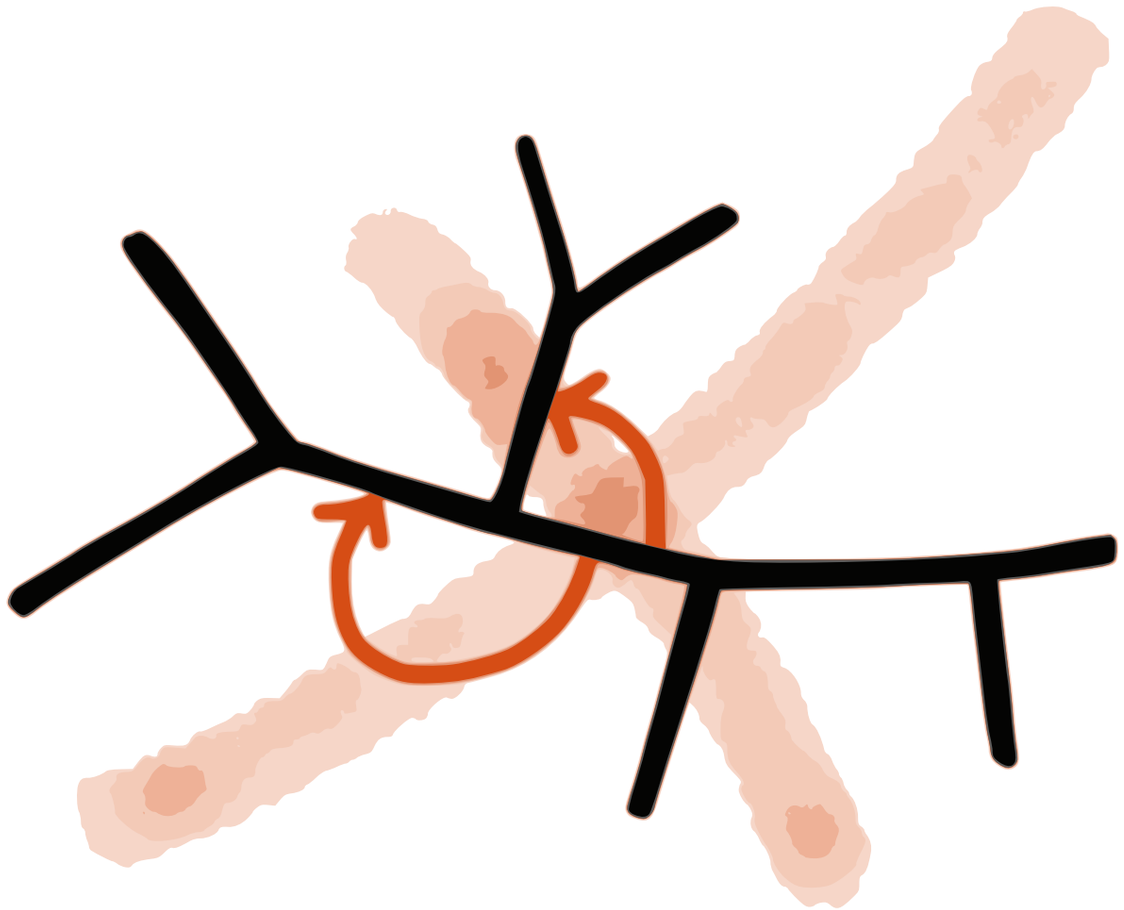
Missing: distance function
Hardwired-cluster distance only for rooted networks
Summary of networks: clades!

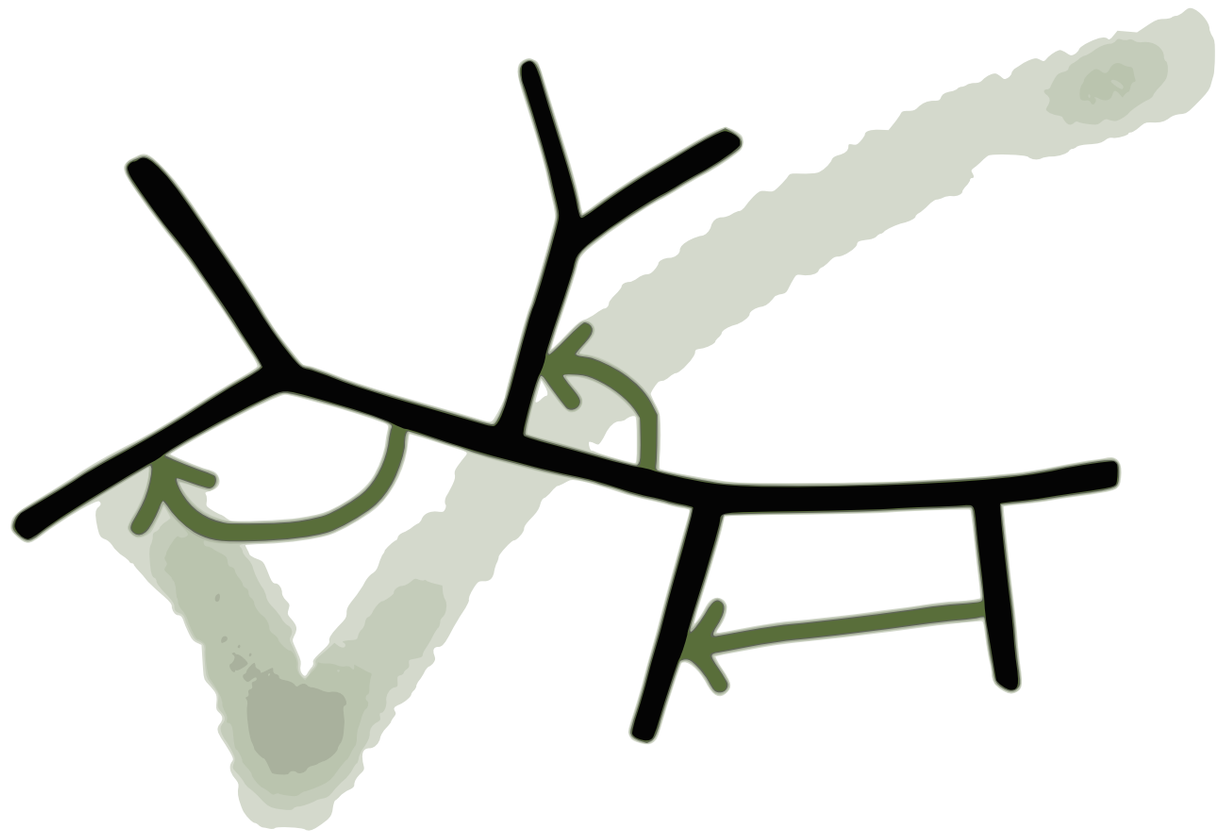
Network summary



(S.-L. et al, 2017, MBE)

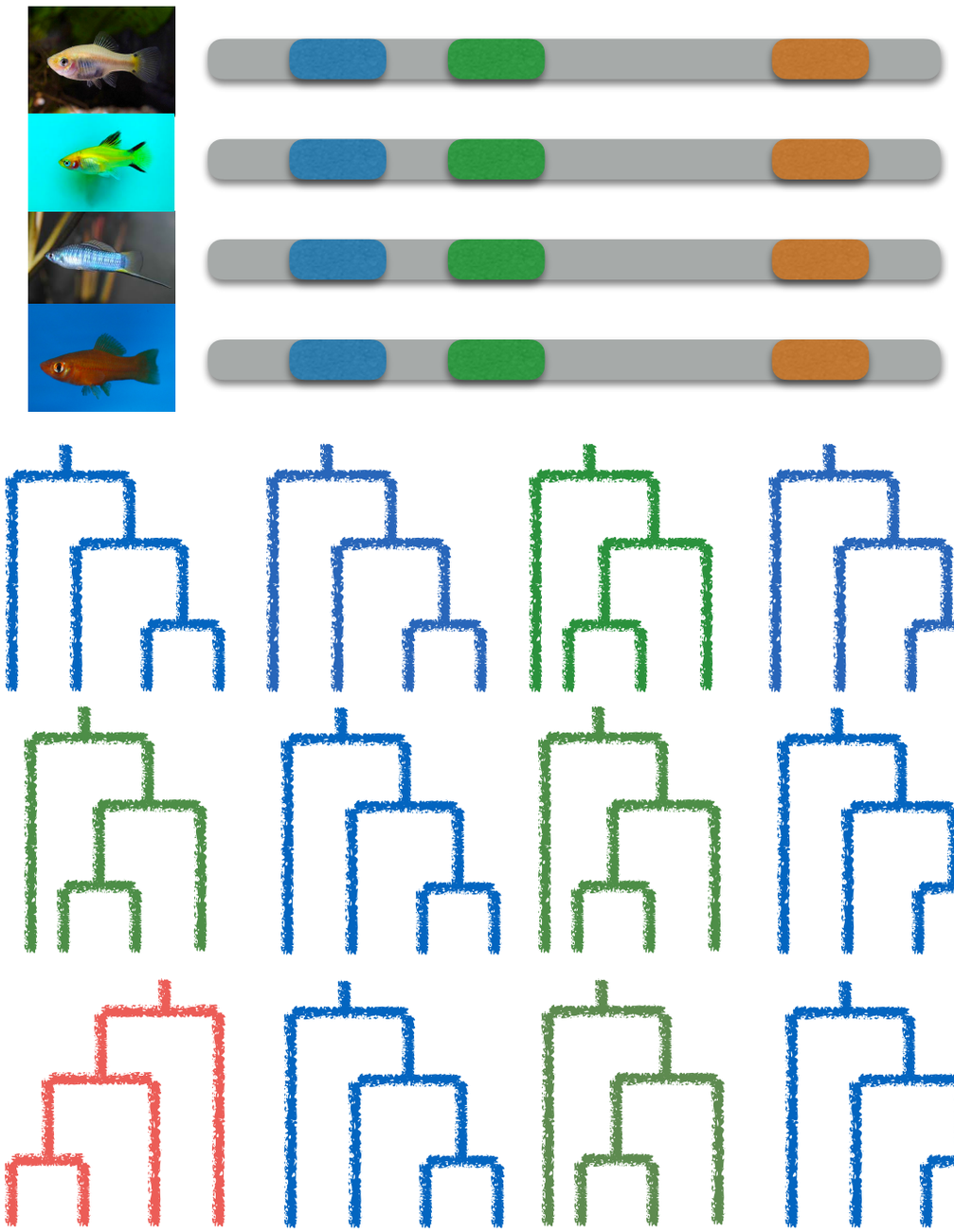
snaQ limitation: Level-1 networks





When?

Phylogenetic network



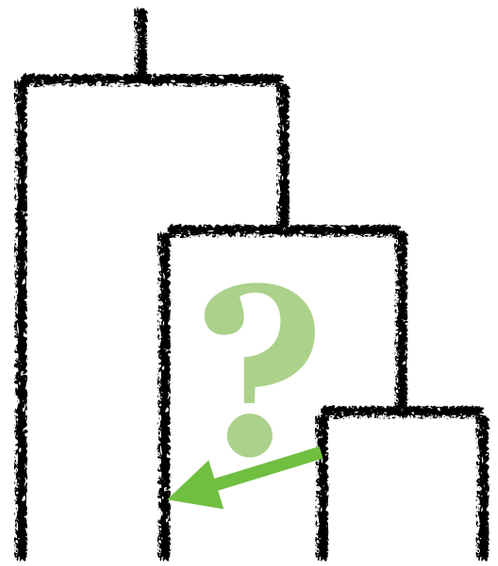
Data

TICR

GitHub

Goodness-of-fit test

Hypothesis test:
Is a tree a good fit?



`https://github.com/nstenz/TICR`

(Stenz et al, 2015, Syst Bio)

PhyloNetworks: analysis for phylogenetic networks

build passing | docs stable | docs dev | codecov 81% | coverage 67%

Overview

PhyloNetworks is a Julia package with utilities to:

- read / write phylogenetic trees and networks, in (extended) Newick format. Networks are considered explicit: nodes represent ancestral species. They can be rooted or unrooted.
- manipulate networks: re-root, prune taxa, remove hybrid edges, extract the major tree from a network, extract displayed networks / trees
- compare networks / trees with dissimilarity measures (Robinson-Foulds distance on trees)
- summarize samples of bootstrap networks (or trees) with edge and node support
- estimate species networks from multilocus data (see below)
- phylogenetic comparative methods for continuous trait evolution on species networks / trees

- Step-by-step tutorial
- Online documentation
- Google user group

(S.-L. et al, 2017, MBE)

 https://solislemuslab.github.io/  @solislemuslab  crsl4

In-class dynamic

- **Time:** 20 minutes

- **Instructions:** We will go over the PhyloNetworks pipeline which will cover ASTRAL, BUCKy and SNaQ. Create our own reproducible script.

- **Options for you:**

1. "I think that I can follow the pipeline by myself or with a small group of peers": you should join the Congregate room

2. "I think I need more one-on-one help to run the commands": you can stay here in the zoom room